Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 863–874
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Kernel PCA for novelty detection

Heiko Hoffmann*,[1]

Max Planck Institute for Human Cognitive and Brain Sciences, Amalienstr. 33, 80799 Munich, Germany

Received 14 November 2005; received in revised form 27 June 2006; accepted 16 July 2006

Abstract

Kernel principal component analysis (kernel PCA) is a non-linear extension of PCA. This study introduces and investigates the use of kernel PCA for novelty detection. Training data are mapped into an infinite-dimensional feature space. In this space, kernel PCA extracts the principal components of the data distribution. The squared distance to the corresponding principal subspace is the measure for novelty. This new method demonstrated a competitive performance on two-dimensional synthetic distributions and on two real-world data sets: handwritten digits and breast-cancer cytology.

Keywords: Kernel method; Novelty detection; PCA; Handwritten digit; Breast cancer

1. Introduction

Novelty detection is one-class classification—for a review see Markou and Singh [1,2]. In training, a machine learns from ordinary data. Later, using previously unknown data, this machine tries to separate ordinary from novel patterns.

One-class classification is useful when normal samples are abundant, but abnormal samples are rare. For example, healthy tissue outweighs malignant cancer. Moreover, if the structure of novel data is obscure, one-class classification might be advantageous over two-class classification. A machine learning technique that works well for data with such characteristics would be of great benefit for medical diagnosis, particularly, for the early diagnoses of cancer, which requires the examination of millions of humans [3,4].

For two-class classification, kernel support-vector machines (kernel SVM) proved to be an excellent choice [5]. This non-linear variant of SVM uses the kernel trick: that is, for algorithms in which the data points occur only within scalar products, the scalar product can be replaced with a kernel function. Thus, the data can be mapped into a higher-dimensional space, a so-called feature space, while the algorithm still operates in the original space (see also Section 2).

Because of the success of SVM, attempts have been made to apply SVM to novelty detection [6–8]. An SVM needs to separate the data against something. In one-class SVM [6,7], the data are separated from the origin in feature space. Alternatively, the support vector domain description (SVDD) [8] encloses the data in feature space with a sphere; for radial-basis-function (RBF) kernels, this procedure gives the same result as the one-class SVM [7]. However elegant, these approaches are not satisfactory because for some training sets, the space enclosed by the corresponding decision boundaries is too large [9].

A different approach to novelty detection is to generate a simplified model of the distribution of training data. For linear distributions, principal component analysis (PCA) is the method of choice, but many interesting distributions are non-linear. In the non-linear case, Gaussian mixture models [10], auto-associative multi-layer perceptrons [11], and principal curves and surfaces [11] have been used. These methods, however, need to solve a non-linear optimization

* Tel.: +44 131 6513437.
E-mail address: mpi@heikohoffmann.de.
[1] Present address: Institute of Perception, Action, and Behavior, School of Informatics, University of Edinburgh, EH9 3JZ, UK.

0031-3203/$30.00
doi:10.1016/j.patcog.2006.07.009

problem and are thus prone to local minima and sensitive to the initialization.

This article combines the distribution-modeling approach with kernel techniques, which do essentially linear algebra. Here, the distribution of training data is modeled by kernel PCA [12], which computes PCA in the feature space. The novelty of the presented approach is to compute the reconstruction error in feature space and to use it as a novelty measure. Decision boundaries herein are iso-potential curves or surfaces of the reconstruction error. Though simple, this method has to my knowledge not been reported before, and it turns out to be a promising novelty detector.

The new method was tested on two-dimensional synthetic distributions and on higher-dimensional real-world data. On the synthetic data, the decision boundaries can follow smoothly the shape of the distribution of data points. To test the performance quantitatively, an ordinary/novel classification task was carried out on two real-world data sets: handwritten digits and breast-cancer cytology. For both data sets, a receiver-operating-characteristic (ROC) analysis [13,14] demonstrates that the new method does better compared with three alternatives: one-class SVM, standard PCA, and the Parzen window density estimator. All of these methods depend on free parameters. However, a proper parameter choice for kernel PCA leads to a performance that cannot be matched by any parameter choice for the three other methods.

The remainder of this article is organized as follows. Section 2 briefly reviews the kernel PCA algorithm. Section 3 describes the extensions to obtain the novelty measure. Section 4 reports the experiments. Section 5 shows a discussion, and Section 6 concludes the article. Appendix A contains a theoretical treatment of the reconstruction error with large kernel widths.

2. Kernel PCA

Kernel PCA [5,12] extends standard PCA to non-linear data distributions. We assume a distribution consisting of n data points $\mathbf{x}_i \in \mathbb{R}^d$. Before performing a PCA, these data points are mapped into a higher-dimensional feature space $\mathcal{F}$,

$$\mathbf{x}_i \to \mathbf{\Phi}(\mathbf{x}_i). \tag{1}$$

In this space, standard PCA is performed. The trick herein is that the PCA can be computed such that the vectors $\mathbf{\Phi}(\mathbf{x}_i)$ appear only within scalar products [12]. Thus, mapping (1) can be omitted. Instead, we only work with a kernel function $k(\mathbf{x}, \mathbf{y})$, which replaces the scalar product $(\mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{y}))$. In kernel PCA, an eigenvector $\mathbf{V}$ of the covariance matrix in $\mathcal{F}$ is a linear combination of points $\mathbf{\Phi}(\mathbf{x}_i)$,

$$\mathbf{V} = \sum_{i=1}^{n} \alpha_i \tilde{\mathbf{\Phi}}(\mathbf{x}_i), \tag{2}$$

with

$$\tilde{\mathbf{\Phi}}(\mathbf{x}_i) = \mathbf{\Phi}(\mathbf{x}_i) - \frac{1}{n}\sum_{r=1}^{n} \mathbf{\Phi}(\mathbf{x}_r). \tag{3}$$

The vectors $\tilde{\mathbf{\Phi}}(\mathbf{x}_i)$ are chosen such that they are centered around the origin in $\mathcal{F}$. The α_i are the components of a vector $\boldsymbol{\alpha}$. It turns out that this vector is an eigenvector of the matrix $\tilde{K}_{ij} = (\tilde{\mathbf{\Phi}}(\mathbf{x}_i) \cdot \tilde{\mathbf{\Phi}}(\mathbf{x}_j))$. The length of $\boldsymbol{\alpha}$ is chosen such that the principal components $\mathbf{V}$ have unit length: $\|\mathbf{V}\| = 1 \Leftrightarrow \|\boldsymbol{\alpha}\|^2 = 1/\lambda$, with λ being the eigenvalue of $\tilde{K}$ corresponding to $\boldsymbol{\alpha}$. To compute $\tilde{K}$, we substitute $\tilde{\mathbf{\Phi}}$ according to Eq. (3). This substitution gives $\tilde{K}_{ij}$ as a function of the kernel matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$:

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}\sum_{r=1}^{n} K_{ir} - \frac{1}{n}\sum_{r=1}^{n} K_{rj} + \frac{1}{n^2}\sum_{r,s=1}^{n} K_{rs}. \tag{4}$$

3. Measure for novelty

This section, first, motivates the reconstruction error in feature space, second, considers the simplified case of computing only the distance to the center of the training data in $\mathcal{F}$, and, third, shows the computation of the reconstruction error.

3.1. Motivation

The new method is geometrically motivated and aims at giving lower classification errors than the one-class SVM. This article focuses on the RBF kernel, particularly, the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2))$, since this kernel is the most common for both one-class SVM and SVDD, and experiments show that for these methods, the Gaussian kernel is more suitable than the polynomial kernel [7–9]. This section illustrates that for RBF kernels, the reconstruction-error decision boundary in $\mathcal{F}$ encloses data in general tighter than the one-class SVM and gives thus a better description of the data (Fig. 1).

For RBF kernels, $k(\mathbf{x}, \mathbf{x})$ takes the same constant value for all $\mathbf{x}$. Therefore, in $\mathcal{F}$, all $\mathbf{\Phi}(\mathbf{x})$ lie on a hyper-dimensional sphere $\mathcal{S}$. Fig. 1 shows only three dimensions of $\mathcal{F}$, but for RBF kernels, $\mathcal{F}$ is infinite-dimensional [5]. However, this illustration is still meaningful since n data points $\mathbf{\Phi}(\mathbf{x}_i)$ can span only a finite space $\mathcal{U}$, which is maximally n-dimensional if we include the origin in $\mathcal{F}$. Due to the rotational invariance of the Euclidean norm, also in $\mathcal{U}$, the data lie on a sphere that is embedded in $\mathcal{U}$ and centered at the origin.

If we require that all data points are enclosed by the decision boundary (hard margins), the SVDD encloses all data points in $\mathcal{F}$ with a sphere as tight as possible, and the one-class SVM puts a plane as close as possible to $\{\mathbf{\Phi}(\mathbf{x}_i)\}$ to separate them from the origin. Since the intersection of the SVDD sphere with $\mathcal{S}$ equals the intersection of the SVM

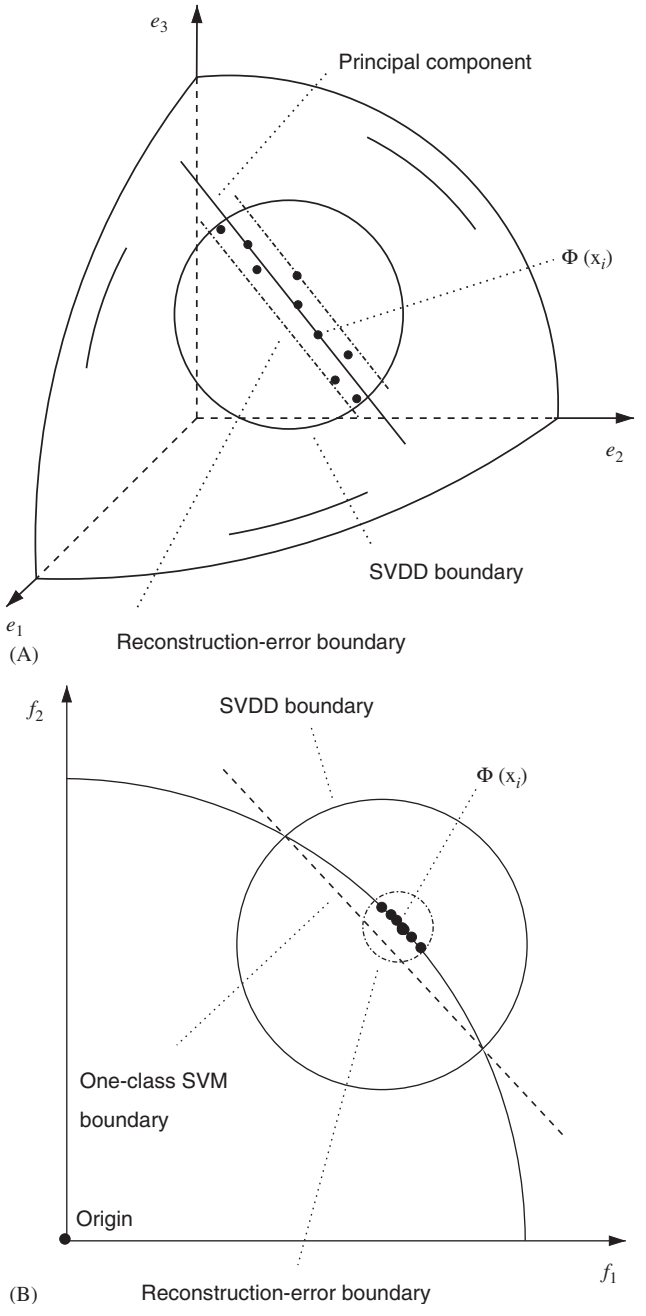

Fig. 1. Decision boundaries in the feature space of an RBF kernel, comparing one-class SVM, SVDD, and the reconstruction error: (A) The boundaries are illustrated in a three-dimensional feature space. All data points $\mathbf{\Phi}(\mathbf{x}_i)$ lie on a sphere. (B) Cross-section through the center of the SVDD sphere and orthogonal to the principal component for the situation in A.

plane with $\mathscr{S}$, the boundary on $\mathscr{S}$ is the same for both methods [7] (a circle for a three-dimensional $\mathscr{U}$, see Fig. 1). This boundary, however, does not tightly enclose the data distribution if the data have a multiform variance in $\mathscr{F}$.

In contrast, the reconstruction error takes into account the heterogeneous variance of the distribution in $\mathscr{F}$. For multi-variate data, orthogonal to the principal subspace, the decision boundary is closer to the data distribution than the SVDD sphere (Fig.1). In the direction of the principal subspace, also a boundary emerges since $\mathscr{S}$ is bending away from the principal subspace (as for the one-class-SVM plane—see Figs. 1B and 5). This emerging boundary ensures that the total boundary is closed; this characteristic seems to be missing for polynomial kernels (Fig. 2, left), where $\{\mathbf{\Phi}(\mathbf{x}_i)\}$ is not restricted to a sphere. To conclude, compared with the one-class SVM, for the same number of enclosed data points, the reconstruction-error boundary encloses a smaller volume in $\mathscr{S}$.

This illustration already gives an insight for choosing the two free parameters of kernel PCA: the kernel width σ and the number of eigenvectors q. The width σ must be within a range of optimal values. For small σ, $k(\mathbf{x}_i, \mathbf{x}_j) \approx 0$ for all i and j with $i \neq j$. Thus, all $\mathbf{\Phi}(\mathbf{x}_i)$ are (almost) orthogonal to each other, and a PCA becomes meaningless. For large σ, the reconstruction error in $\mathscr{F}$ approaches the reconstruction error for standard PCA (see Appendix A). Furthermore, q needs to be sufficiently large, because otherwise, the reconstruction error is high for some points within the data distribution. Consequently, the threshold on the novelty measure would be also high leading to a loose decision boundary (the same holds also for standard PCA). The dependence on σ and q is studied in the experimental section.

3.2. Spherical potential

With no principal components, the reconstruction error reduces to a spherical potential field in feature space. All we need is the center of the data in $\mathscr{F}$, $\mathbf{\Phi}_0 = 1/n \sum_{i=1}^{n} \mathbf{\Phi}(\mathbf{x}_i)$. The potential of a point $\mathbf{z}$ in the original space is the squared distance from the mapping $\mathbf{\Phi}(\mathbf{z})$ to the center $\mathbf{\Phi}_0$,

$$p_S(\mathbf{z}) = \|\mathbf{\Phi}(\mathbf{z}) - \mathbf{\Phi}_0\|^2. \tag{5}$$

The squared magnitude can be written with kernel functions using the above expression for $\mathbf{\Phi}_0$,

$$p_S(\mathbf{z}) = k(\mathbf{z}, \mathbf{z}) - \frac{2}{n} \sum_{i=1}^{n} k(\mathbf{z}, \mathbf{x}_i) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j). \tag{6}$$

All parts of this equation are known. The last term is constant, and can therefore be omitted. For RBF kernels, the first term is also constant, and the potential can be simplified to

$$\tilde{p}_S(\mathbf{z}) = -\frac{2}{n} \sum_{i=1}^{n} k(\mathbf{z}, \mathbf{x}_i). \tag{7}$$

This function is up to a multiplicative constant equal to the Parzen window density estimator [15].

3.3. Reconstruction error

As novelty measure, we use the reconstruction error [16] in feature space

$$p(\tilde{\mathbf{\Phi}}) = (\tilde{\mathbf{\Phi}} \cdot \tilde{\mathbf{\Phi}}) - (W\tilde{\mathbf{\Phi}} \cdot W\tilde{\mathbf{\Phi}}). \tag{8}$$

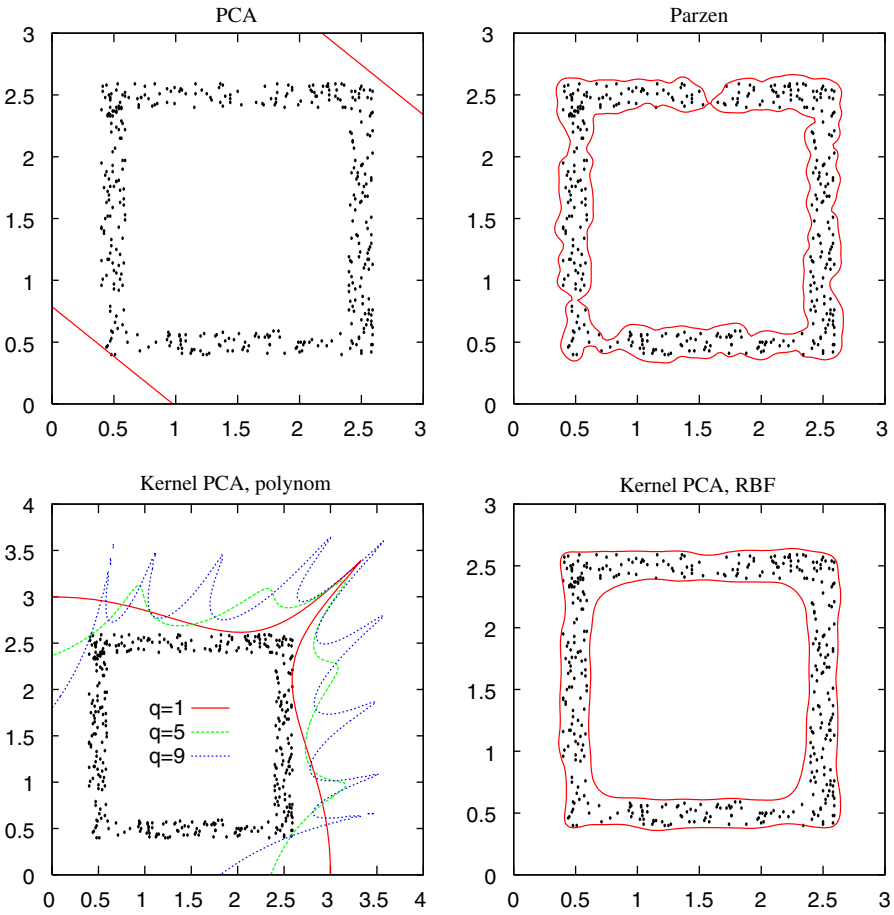

Fig. 2. Decision boundaries for various methods: (top left) PCA reconstruction error with $q=1$ eigenvector, (top right) Parzen window density estimator with $\sigma=0.05$, (bottom left) reconstruction error in $\mathcal{F}$ with polynomial kernel $k(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i\cdot\mathbf{x}_j)^{10}$ and various q values, (bottom right) reconstruction error in $\mathcal{F}$ with Gaussian kernel using $\sigma=0.4$ and $q=40$.

$\tilde{\mathbf{\Phi}}$ is a vector originating from the center of the distribution in feature space, $\tilde{\mathbf{\Phi}}(\mathbf{z})=\mathbf{\Phi}(\mathbf{z})-\mathbf{\Phi}_0$. Let q be the number of principal components. The matrix W contains the q row vectors $\mathbf{V}^l$. The index l denotes the lth eigenvector, with $l=1$ for the eigenvector with the largest eigenvalue.

We need to eliminate $\tilde{\mathbf{\Phi}}$ in Eq. (8), and write the potential as a function of a vector $\mathbf{z}$ taken from the original space. The projection $f_l(\mathbf{z})$ of $\tilde{\mathbf{\Phi}}(\mathbf{z})$ onto the eigenvector $\mathbf{V}^l=\sum_{i=1}^n \alpha_i^l\tilde{\mathbf{\Phi}}(\mathbf{x}_i)$ can be readily evaluated using the kernel function k,

$$\begin{aligned} f_l(\mathbf{z}) = (\tilde{\mathbf{\Phi}}(\mathbf{z})\cdot\mathbf{V}^l) &= \left(\left[\mathbf{\Phi}(\mathbf{z})-\frac{1}{n}\sum_{r=1}^{n}\mathbf{\Phi}(\mathbf{x}_r)\right]\right. \\ &\quad \left.\cdot\left[\sum_{i=1}^{n}\alpha_i^l\mathbf{\Phi}(\mathbf{x}_i)-\frac{1}{n}\sum_{i,r=1}^{n}\alpha_i^l\mathbf{\Phi}(\mathbf{x}_r)\right]\right) \\ &= \sum_{i=1}^{n}\alpha_i^l\left[k(\mathbf{z},\mathbf{x}_i)-\frac{1}{n}\sum_{r=1}^{n}k(\mathbf{x}_i,\mathbf{x}_r)\right. \\ &\quad \left.-\frac{1}{n}\sum_{r=1}^{n}k(\mathbf{z},\mathbf{x}_r)+\frac{1}{n^2}\sum_{r,s=1}^{n}k(\mathbf{x}_r,\mathbf{x}_s)\right]. \end{aligned} \tag{9}$$

Here, the second equality uses Eq. (3). As a result, $p(\tilde{\mathbf{\Phi}})$ can be expressed as

$$p(\tilde{\mathbf{\Phi}})=(\tilde{\mathbf{\Phi}}\cdot\tilde{\mathbf{\Phi}})-\sum_{l=1}^{q}f_l(\mathbf{z})^2. \tag{10}$$

The scalar product $(\tilde{\mathbf{\Phi}}\cdot\tilde{\mathbf{\Phi}})$ equals the spherical potential (6). Thus, the expression of the potential $p(\mathbf{z})$ can be further simplified

$$p(\mathbf{z})=p_S(\mathbf{z})-\sum_{l=1}^{q}f_l(\mathbf{z})^2. \tag{11}$$

This is the desired form of the novelty measure in $\mathbb{R}^d$.

The above computation of $f_l(\mathbf{z})$ requires n evaluations of the kernel function for each $\mathbf{z}$. Since for all l components, the same kernels can be used, the total number of kernel evaluations is also n.

4. Experiments

The decision boundaries for the new method are illustrated using two-dimensional synthetic distributions. Furthermore,

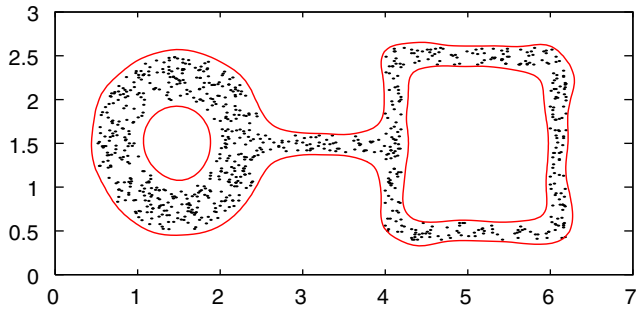

Fig. 3. Decision boundary for the ring-line-square distribution using the reconstruction error in $\mathcal{F}$ with $\sigma = 0.4$ and $q = 40$.

this method is applied to higher-dimensional data, handwritten digits and breast-cancer cytology.

4.1. Methods

The methods section comprises the different data sets, the implementation and evaluation of kernel PCA, and the alternative novelty detectors used for comparison.

4.1.1. Data sets

Kernel PCA for novelty detection was tested on synthetic and real-world data sets. Five synthetic distributions were used: square, square-noise, ring-line-square, spiral, and sine-noise.

Square: The square consists of four lines, 2.2 long and 0.2 wide (Fig. 2). Within the area of these lines, 400 points were randomly distributed with equal probability.

Square-noise: The square-noise was generated by adding to the above distribution 50 noise points randomly drawn from the area $\{(x, y)|x \in [0, 3],\ y \in [0, 3]\}$, which surrounds the square (Fig. 5).

Ring-line-square: The ring-line-square distribution is composed of a ring with an inner diameter of 1.0 and an outer diameter of 2.0, a square with the size as described above, and a 1.6 long and 0.2 wide line connecting the two parts (Fig. 3). Within the area of these three parts, 850 points were randomly distributed with equal probability.

Spiral: The area of the spiral is defined by the set $\{(x, y)\,|\,x = (0.07\varphi + a)\cos(\varphi),\ y = (0.07\varphi + a)\sin(\varphi), a \in [0, 0.1], \varphi \in [0, 6\pi]\}$. Within this area, 700 points were randomly distributed with equal probability (Fig. 4).

Sine-noise: The sine-noise distribution consists of a sine-wave and surrounding noise (Fig. 6). In the sine-wave part, 500 points are uniformly distributed along $y = 0.8\sin(2\varphi)$ with $\varphi \in [0, 2\pi]$. These points are surrounded by 200 points that were distributed randomly with equal probability in the rectangle $\{(x, y)|x \in [0, 2\pi],\ y \in [-1.5, 1.5]\}$.

Two real-world data sets were used: handwritten digits and breast-cancer data.

Digit 0: The digits were obtained from the MNIST digit database [17]. The original 28×28 pixels images are almost binary (see Fig. 11). Thus, the digits occupy only the corners of a 784-dimensional hyper-cube. To get a more continuous distribution of digits, the original images were blurred and sub-sampled down to 8×8 pixels. The MNIST database is split into training set and test set. To train the novelty detectors, the first 2 000 '0' digits from the training set were used. For the 0/not-0 classification task, from the test set, all 980 '0' digits were used together with the first 109 samples from each other digit.

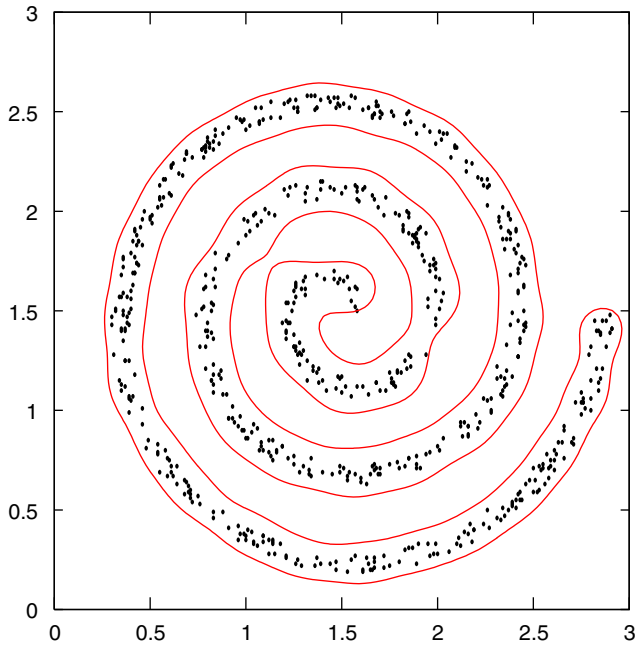

Fig. 4. Decision boundary for the spiral distribution using the reconstruction error in $\mathcal{F}$ with $\sigma = 0.25$ and $q = 40$.

Cancer: The breast-cancer data were obtained from the UCI machine-learning repository [18]. These data were collected by Dr. William H. Wolberg at the University of Wisconsin Hospitals in Madison [19]. The patterns in this data set belong to two classes: *benign* and *malignant*. Each pattern consists of nine cytological characteristics such as, for example, the uniformity of cell size. Each of these characteristics is graded with an integer value from 1 to 10, with 1 being typical benign. The database contains some patterns with missing attributes, these patterns were removed before further processing. The remaining patterns were scaled to have unit variance in each dimension. To avoid numerical errors because of the discrete values, a uniform noise from the interval $[-0.05, 0.05]$ was added to each value. The novelty detectors were trained on the first 200 benign samples. The remaining samples were used for testing: 244 benign and 239 malignant.

4.1.2. Kernel PCA implementation and evaluation

Kernel PCA was computed on all data points of each distribution. Unless otherwise noted, a Gaussian kernel with width σ was used. Exploratory tests with two other RBF

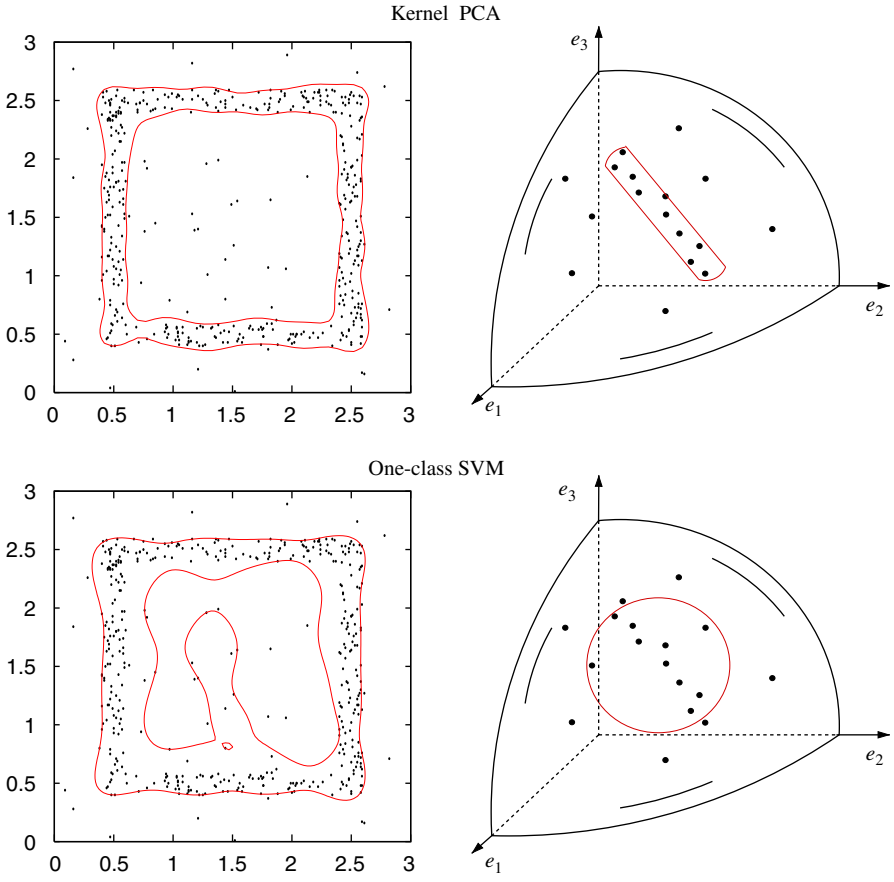

Fig. 5. Decision boundaries on noisy data comparing kernel PCA ($\sigma = 0.3$, $q = 20$) with the one-class SVM ($\sigma = 0.362$, $\nu = 1/9$). (Left) Result on the square-noise set. (Right) Illustration of the corresponding decision boundaries in $\mathcal{F}$ (see Fig. 1).

kernels, the 'multi-quadratic' and the 'inverse multi-quadratic' [5], gave similar optimal results. However, since the Gaussian kernel is commonly used for the one-class SVM, the SVDD, and the Parzen density, this kernel is the only RBF kernel presented here.

The eigenvectors $\boldsymbol{\alpha}$ of $\tilde{K}$ were extracted using the routine 'dsyevx' from the linear algebra package LAPACK. This package is based on BLAS, a standard for basic linear algebra computations. Both BLAS and LAPACK are Fortran77 libraries, which can be linked into C/ C++ code. Here, BLAS was optimized for the Athlon C++ using the software ATLAS, version 3.6.0 (see http://math-atlas.sourceforge.net/).

The classification performance on the real-world data was evaluated using ROC curves. An ROC curve plots the fraction of test patterns correctly classified as novel (true positives) versus the fraction of patterns incorrectly classified as novel (false positives) to illustrate the performance over all possible decision thresholds (see Fig. 8). To compute such a curve, first, the reconstruction error $p(\mathbf{z}_i)$ was evaluated for all test patterns i. Second, the set $\{p(\mathbf{z}_i)\}$ was sorted according to the p-values. Finally, by counting how many novel and ordinary samples are above a decision threshold taken between two neighboring p-values, the fractions of true and false positives are readily available. Thus, for each $p(\mathbf{z}_i)$, there is a point on the ROC curve. Together, these points cover the full range of false positives: from 0 to 1.

4.1.3. Alternative novelty detectors

Three alternatives to kernel PCA were tested: the Parzen window density estimator, standard PCA, and the one-class SVM [6] (SVDD produces the same result as the one-class SVM). Thus, the comparison is limited to those methods that are related to the new method. Excluded are methods, like Gaussian mixture models, that require iterative solvers and many parameters, since the consequent multitude of possible outcomes makes a just comparison seemingly impossible (a comparison as in Figs. 9 and 10 would be impossible).

The Parzen window density estimator constructs a probability density function (pdf) from a data distribution by summing kernel functions centered around each data point [15]. This pdf is therefore proportional to the spherical potential in feature space (7), see Section 3.2. For all experiments, a Gaussian kernel with width σ was chosen. On the square data, the optimal σ was chosen to maximize the mean probability across data points if each point's probability is

computed based on the pdf given all other data points (leave-one-out cross-validation).

PCA uses the reconstruction error in $\mathbb{R}^d$ as novelty measure. Thus, the PCA detector was obtained as a special case of the new method by choosing $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})$.

The one-class SVM was tested using the library `LIBSVM`, version 2.71 [20]. A Gaussian kernel with width σ was used for all experiments. For the synthetic distributions that include noise, the value of σ was chosen to maximize the number of noisy points outside the decision boundary plus the number of regular points inside the boundary. To adjust the decision threshold, one-class SVM has a further parameter ν. Its value is approximately equal to the fraction of false positives in a novelty classification task [6]. Thus, the ROC curves were computed by varying ν.

4.2. Results

This section compares qualitatively kernel PCA for novelty detection with other methods, studies the influence of noise in the training data on the decision boundary, investigates the dependence on the kernel parameter σ and on the number of eigenvectors q, and finally, based on the digit 0 and the cancer set compares quantitatively kernel PCA with other methods.

4.2.1. Decision boundaries on synthetic data

On the square data set, Fig. 2 compares qualitatively the decision boundaries for PCA, the Parzen density, kernel PCA with a polynomial kernel, and kernel PCA with a Gaussian kernel (here, the one-class SVM was omitted since its result was similar to the one obtained with kernel PCA using a Gaussian kernel). The decision thresholds were chosen, such that the boundaries enclose all training points as tight as possible.

The Parzen density does not generalize well: the decision boundary follows the irregularities of the distribution (a larger σ did not improve the boundary either—see Fig. 7). A linear model like PCA cannot describe the square distribution. Kernel PCA with a polynomial kernel does not suitably describe this distribution either (other polynomial degrees did not give significantly better results). With a Gaussian kernel, however, the decision boundary follows the shape of the distribution without getting disturbed by local irregularities. This ability is further illustrated using the ring-line-square (Fig. 3) and the spiral distribution (Fig. 4).

To test how the reconstruction-error boundaries cope with noise within the training data, the square-noise and the sine-noise distributions were used, and the result is compared with the one-class SVM (Figs. 5 and 6). For kernel PCA, the decision threshold was chosen such that the fraction of outliers equals the given fraction η of noise points; for the one-class SVM, ν was set equal to η. Using the reconstruction error in $\mathcal{F}$, the decision boundary can enclose smoothly the main part of the data, almost undisturbed by noise;[2] the SVM boundary extends into the outlier region.

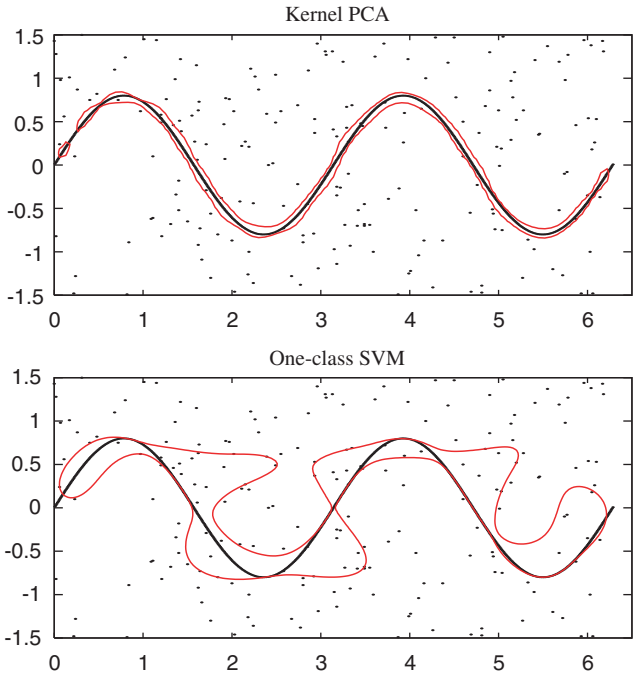

Fig. 6. Decision boundaries for the sine-noise distribution comparing kernel PCA ($\sigma = 0.4$, $q = 40$) with the one-class SVM ($\sigma = 0.489$, $\nu = \frac{2}{7}$).

Kernel PCA depends on the number of principal components q and on the width of the Gaussian kernel σ. On the square distribution, for small σ, increasing the number of eigenvectors changed only little the shape of the decision boundary (Fig. 7). Increasing both σ and q resulted in a good performance (Fig. 7).

4.2.2. Novelty detection on real-world data

On both the digit 0 and the cancer data set, kernel PCA for novelty detection could achieve lower classification errors compared with the one-class SVM, PCA, and the Parzen density (Figs. 8–10). Since all tested methods depend on free parameters, the area under the ROC curve is shown as a function of these parameters. This illustration demonstrates that for PCA, the Parzen density, and the one-class SVM, no parameter choice is possible to match the optimal performance of the kernel PCA method.

In the kernel PCA and Parzen case, σ has a lower limit (Fig. 9), because below this limit, samples exist from both ordinary and novel classes that numerically reach the maximal potential p (due to the rapidly decreasing Gaussian function). Thus, tests for these low σ values were omitted. For the one-class SVM, the range of possible σ values was even

[2] The number of principal components is lower compared with the no-noise case; otherwise, the boundary would also include nearby noise.

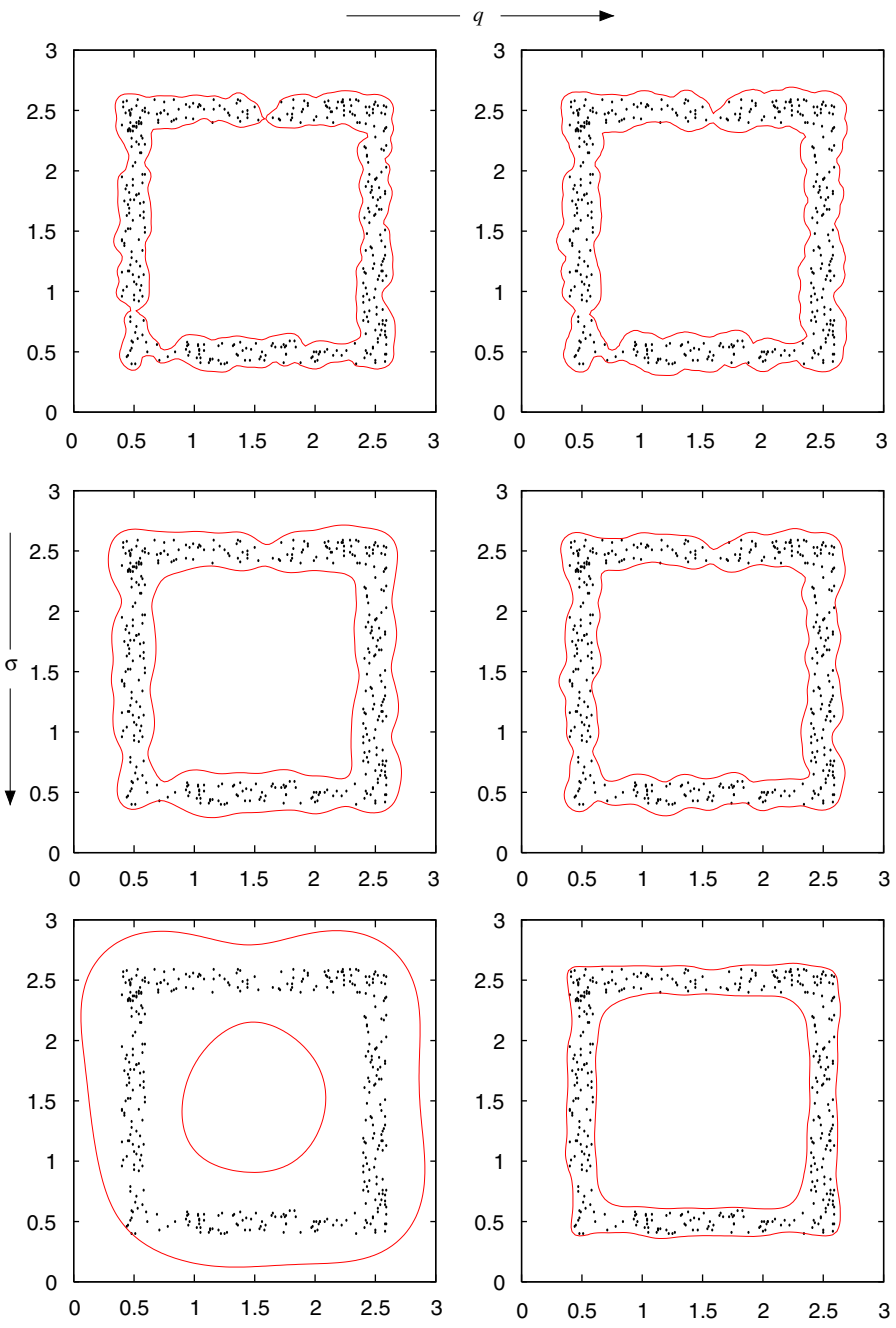

Fig. 7. Dependence on the kernel width σ and on the number of eigenvectors q. The width σ increases from top to down, taking the three values: 0.05, 0.1 and 0.4. The number q increases from left ($q = 0$) to right ($q = 40$).

smaller. Using a small σ, for any choice of ν, the number of false positives did not fall below a limit (Fig. 8). Therefore, the area under such an ROC curve could not be computed. However, also these ROC curves are well below the optimal curve from kernel PCA (Fig. 8).

The new kernel PCA method achieved high ROC areas on both data sets (maximum areas of 0.9953 and 0.9971 for digit and cancer, respectively), even though, the structure of the data appears to be different in both cases. On the digit set, a linear model like PCA does better than the Parzen density (maximum area: 0.9893 versus 0.9873); on the cancer set, however, the Parzen density does very good (maximum area: 0.9966), but PCA does much worse (maximum area: 0.9828).

Furthermore, the results show the behavior in the limit of small and large σ values. For small σ, the performance of the reconstruction error in $\mathcal{F}$ and of the Parzen density become almost equal (Fig. 9). This behavior matches

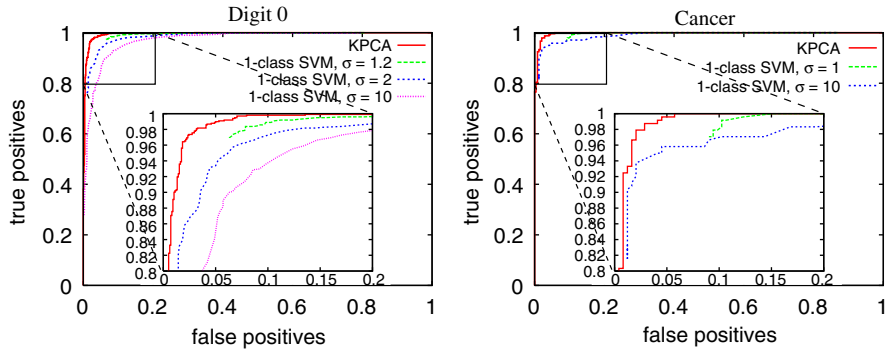

Fig. 8. ROC curves on the digit and the cancer data. Kernel PCA using $\sigma = 4$ and $q = 100$ (digit 0) and $\sigma = 2$ and $q = 190$ (cancer) is compared with the one-class SVM for various σ values.

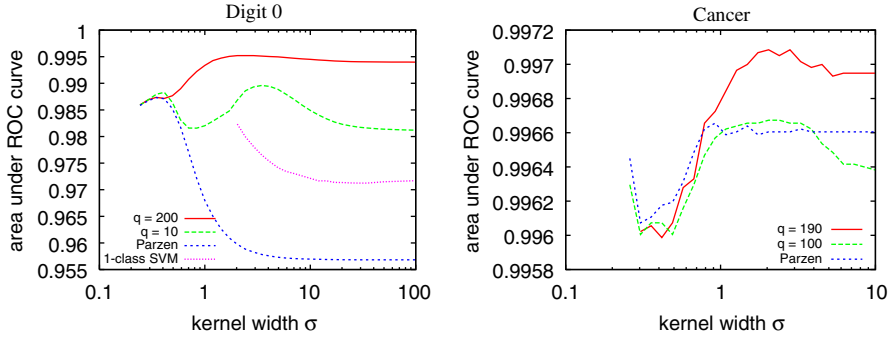

Fig. 9. Performance depending on the kernel width σ using the digit and the cancer data. Kernel PCA for various q values is compared with the Parzen density and the one-class SVM. For the cancer data, the ROC area of the one-class SVM is below the range shown in this diagram.

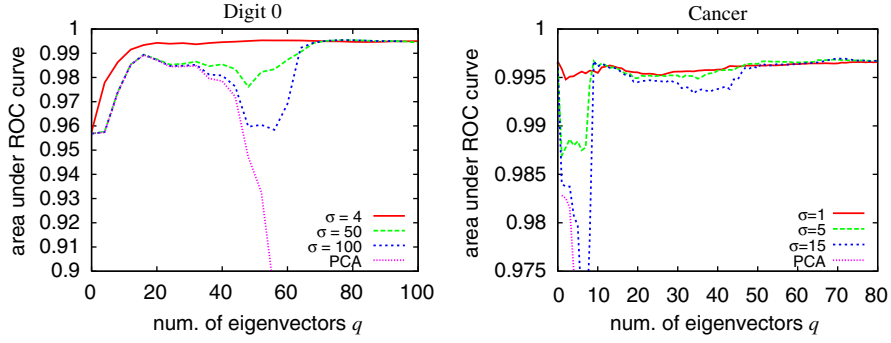

Fig. 10. Performance depending on the number of eigenvectors q using the digit and the cancer data. Kernel PCA for various σ values is compared with standard PCA.

the observation of the decision boundaries (Fig. 7). For large σ, the performance reaches the level of PCA, if q is smaller than d—otherwise PCA is meaningless (Fig. 10). This limit behavior can be also predicted theoretically (see Appendix A).

A final test further illustrates the proper function of the new method. The reconstruction error in $\mathcal{F}$ was used to find unusual '0' digits within the MNIST test set. Fig. 11 displays the 10 digits that had the highest reconstruction errors. Most of these samples look indeed unusual.

5. Discussion

This section discusses the effect of noise in the training data, mentions concerns about the computational complexity of the new method, and points out other related methods.

5.1. Noisy data

Novelty detectors learn from data that are assumed to contain only representatives of the ordinary class. In real applications, however, these data contain noise, thus, outliers.

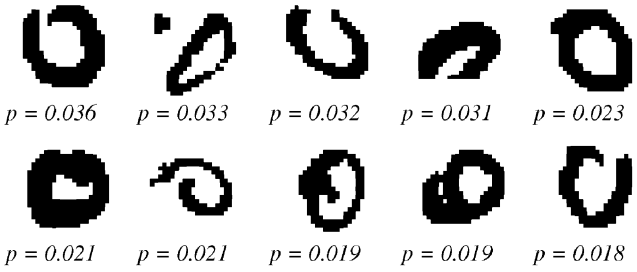

Fig. 11. The 10 most unusual '0' digits from the MNIST test set. The digits are arranged in descending order of their reconstruction error p ($\sigma=4$, $q=100$). The figure shows the unprocessed digits of size 28×28 pixels; for novelty detection, however, the processed digits (8×8 pixels) were used.

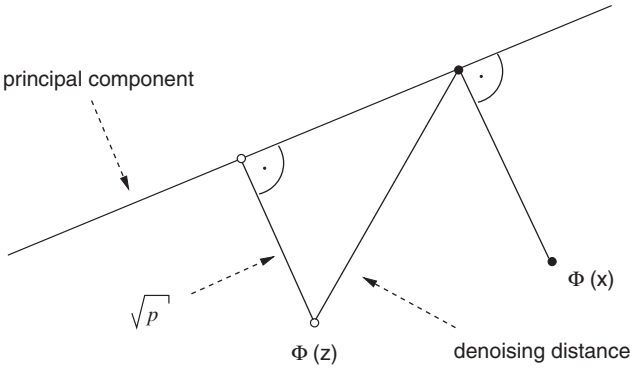

Fig. 12. The difference between the distance to be optimized in denoising and the reconstruction error p.

Therefore, a good novelty detector should be robust to noise. In the presented model, outliers might distort the principal components extracted by kernel PCA. Like PCA, its kernel variant is not robust against such noise [21,22]. However, as for PCA, also for kernel PCA, robust versions exist [21,22]. These approaches essentially try to remove outliers, either before or in alternation with computing the principal components. The presented method has the advantage that it uses a standard algorithm, kernel PCA. Thus, improvements and modifications of this algorithm can be readily applied.

No robust versions of kernel PCA were used here for the reported experiments; nevertheless, the new decision boundaries appeared to be robust under noise (Figs. 5 and 6). This robustness may be explained by the almost uniform variance of the added noise; thus, probably, the principal components were undisturbed. This noise, however, did disturb the results of the one-class SVM (Figs. 5 and 6). In feature space, the sphere (the SVM boundary on $\mathscr{S}$) that encloses the data is less tight around the noise-free part of the distribution than the reconstruction-error boundary (Section 3.1). Thus, for the same number of enclosed points, the one-class SVM encloses more outliers than the new method (Fig. 5, right).

5.2. Computational complexity

Kernel PCA is computationally expensive. Most time-consuming is the extraction of the eigenvectors of $\tilde{K}$, which is $O(n^3)$ if extracting all eigenvectors [23]. Additionally, if searching for the parameters σ and q, for example, by cross-validation, the computation time is further multiplied by the number of PCA evaluations. Kernel PCA is also memory exhaustive: the $n\times n$ matrix $\tilde{K}$ needs to be stored, and in the presented experiments, only a tiny fraction of entries in $\tilde{K}$ were almost zero. Therefore, on large data sets, Monte-Carlo sampling is necessary.

Furthermore, testing is expensive. For each new data point, the kernel function needs to be evaluated n-times. However, this number could be reduced using so-called 'reduced-set methods' [24,25].

The one-class SVM is faster: using an Athlon 1800+, on the digit-0 data, with $\sigma=2$ and $\nu=0.1$, `LIBSVM` needed 1.3 s for training and 0.5 s to classify all test patterns. In contrast, kernel PCA with $\sigma=4$ and $q=100$ needed 31.6 s for training (computing the kernel matrix and extracting eigenvectors) and 34.4 s for testing. However, computing an ROC curve using the reconstruction error does not require any noticeable additional time, but the one-class SVM needs to be retrained for different ν values.

5.3. Related methods

The reconstruction error in $\mathscr{F}$ is similar to but differs from the squared error in the denoising application of kernel PCA [26]. To denoise a pattern $\mathbf{x}$, its mapping $\mathbf{\Phi}(\mathbf{x})$ is projected onto the principal subspace. The denoised pattern $\mathbf{z}$ is obtained by minimizing the squared distance between $\mathbf{\Phi}(\mathbf{z})$ and the projection of $\mathbf{\Phi}(\mathbf{x})$. Fig. 12 illustrates the difference to the reconstruction error.

Tax and Juszczak [9] used kernel PCA as a preprocessing step for novelty detection. Kernel PCA is used to whiten (to make the variance in each direction equal) the data in feature space. Later, these data are enclosed by a sphere to obtain a one-class classifier [8]. A problem for whitening are directions with variance close to zero. In the present study, the variance is close to zero for most directions a distribution expands to (if σ is not too small). Thus, whitening is disadvantageous because on the one hand, normalizing these directions is prone to computational errors, and on the other hand, omitting these directions ignores the low variance, since everything will be enclosed in a sphere.

6. Conclusions

This article studied kernel PCA for novelty detection. A principal subspace in an infinite-dimensional feature space described the distribution of training data. The reconstruction error of a new data point with respect to this

subspace was used as a measure for novelty. This new method demonstrated a higher ordinary/novel-classification performance on a handwritten-digit and a breast-cancer database compared with the one-class SVM and the Parzen window density estimator. Both of these methods were competitive in past experiments [1,2].

Using the reconstruction error in feature space, the decision boundaries followed smoothly the shape of two-dimensional synthetic distributions, without getting distorted by the position of single data points. Thus, the new method appears to generalize better compared with the Parzen density. Furthermore, compared with the one-class SVM, the presented method demonstrated to be more robust against noise within the training set.

This article demonstrated the dependence on the kernel parameter σ and on the number of eigenvectors q. For small σ, the new method behaved like the Parzen density. For large σ, the reconstruction error in feature space approaches the reconstruction error for standard PCA. The number of eigenvectors q had to be sufficiently large for a near optimal performance on both real-world data sets and on the synthetic distributions without noise.

Future work aims at finding rules for choosing the two free parameters, σ and q (without the need of a time-consuming cross-validation step). Of advantage should be therein that a wide range of parameters resulted in a near optimal performance. Furthermore, the range of usable data distributions needs to be explored. The arguments in Section 3.1 already suggest that such distributions have to originate from an underlying manifold or from some locally connected structure, as also most dimension-reduction techniques require (see, for example, mixture of local PCA [27] and locally linear embedding [28]).

Acknowledgments

I thank the anonymous reviewers, whose comments helped to improve the manuscript.

Appendix A. Behavior at large kernel widths

This section analyzes theoretically the behavior of the reconstruction error in feature space at large widths σ of the Gaussian kernel. For $\sigma \gg \max \|\mathbf{x}_i - \mathbf{x}_j\|$ and $q < d$, the 'kernelized' reconstruction error approaches the reconstruction error in the original space $\mathbb{R}^d$.

We assume $\sigma \gg \max \|\mathbf{x}_i - \mathbf{x}_j\|$, and thus, ignore orders smaller than $O(1/\sigma^2)$. Therefore, the Gaussian kernel can be approximated as $k(\mathbf{x}, \mathbf{y}) \approx 1 - \|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2)$. We will show that with this approximation, the reconstruction error (11) is proportional to Eq. (11) with $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})$, which corresponds to standard PCA. First, substituting the approximated kernel into the spherical-potential component (6) gives

$$p_S(\mathbf{z}) \approx \frac{1}{\sigma^2}\left[(\mathbf{z}\cdot\mathbf{z}) - \frac{2}{n}\sum_i (\mathbf{z}\cdot\mathbf{x}_i) + \frac{1}{n^2}\sum_{i,j}(\mathbf{x}_i\cdot\mathbf{x}_j)\right]. \tag{A.1}$$

This potential p_S has two important properties: it scales as $1/\sigma^2$, and the expression in the square brackets equals the spherical-potential component for the polynomial kernel $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})$.

In addition, we use the above substitution to compute the eigenvector projections (9)

$$f_l(\mathbf{z}) \approx \frac{1}{\sigma^2}\sum_i \alpha_i^l \left[(\mathbf{z}\cdot\mathbf{x}_i) - \frac{1}{n}\sum_r (\mathbf{x}_i\cdot\mathbf{x}_r) - \frac{1}{n}\sum_r (\mathbf{z}\cdot\mathbf{x}_r) + \frac{1}{n^2}\sum_{r,s}(\mathbf{x}_r\cdot\mathbf{x}_s)\right]. \tag{A.2}$$

Again, the expression in the square brackets is the same as for $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})$. We still need to evaluate how α_i^l depends on the approximated kernel function. The variables α_i^l are the components of the eigenvectors of the kernel matrix $\tilde{K}$, which is computed according to Eq. (4) such that the data have zero mean in feature space (see Section 2). Substituting the approximated kernel function into Eq. (4) gives

$$\tilde{K}_{ij} \approx \frac{1}{\sigma^2}\left[(\mathbf{x}_i\cdot\mathbf{x}_j) - \frac{1}{n}\sum_r (\mathbf{x}_i\cdot\mathbf{x}_r) - \frac{1}{n}\sum_r (\mathbf{x}_j\cdot\mathbf{x}_r) + \frac{1}{n^2}\sum_{r,s}(\mathbf{x}_r\cdot\mathbf{x}_s)\right]. \tag{A.3}$$

Apart from the factor $1/\sigma^2$, this formula is the same as for $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})$. Thus, in both cases, also the eigenvectors are the same, but the eigenvalues differ by the factor $1/\sigma^2$. Since the length of $\boldsymbol{\alpha}$ is $1/\sqrt{\lambda}$ (see Section 2), the components α_i^l for the approximated Gaussian kernel equal the corresponding components for PCA times σ. Thus, since f_l is squared in Eq. (11), we again have the same $1/\sigma^2$ factor as in Eq. (A.1). In total, the reconstruction error in feature space differs only by a constant factor from the reconstruction error in the original space. Therefore, the results on novelty detection are the same.

References

[1] M. Markou, S. Singh, Novelty detection: a review, part 1: statistical approaches, Signal Process. 83 (12) (2003) 2481–2497.

[2] M. Markou, S. Singh, Novelty detection: a review, part 2: neural network based approaches, Signal Process. 83 (12) (2003) 2499–2521.

[3] L. Tarassenko, P. Hayton, N. Cerneaz, M. Brady, Novelty detection for the identification of masses in mammograms, in: Proceedings of the Fourth IEE International Conference on Artificial Neural Networks, IEE, London, 1995, pp. 442–447.

[4] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai, H. Du, Approaches for automated detection and classification of masses in mammograms, Pattern Recognition 39 (4) (2006) 646–668.
[5] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[6] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, Adv. Neural Inf. Process. Syst. 12 (2000) 582–588.
[7] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 13.
[8] D.M.J. Tax, R.P.W. Duin, Support vector domain description, Pattern Recognition Lett. 20 (1999) 1191–1199.
[9] D.M.J. Tax, P. Juszczak, Kernel whitening for one-class classification, Lecture Notes in Computer Science, vol. 2388, 2002, pp. 40–52.
[10] L. Tarassenko, A. Nairac, N. Townsend, P. Cowley, Novelty detection in jet engines, in: IEE Colloquium on Condition Monitoring: Machinery, External Structures and Health (Ref. No. 1999/034), IEE, London, 1999, pp. 4/1–4/5.
[11] S.O. Song, D. Shin, E.S. Yoon, Analysis of novelty detection properties of autoassociators, in: A.G. Starr, R.B.K.N. Rao (Eds.), Proceedings of the 14th International Congress and Exhibition on Condition Monitoring and Diagnostic Engineering Management (COMADEM), Elsevier Science, Amsterdam, 2001, pp. 577–584.
[12] B. Schölkopf, A.J. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[13] A. Wald, Statistical Decision Functions, Wiley, New York, 1950.
[14] L. Lusted, Introduction to Medical Decision Making, Thomas, Springfield, IL, 1968.
[15] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[16] K.I. Diamantaras, S.Y. Kung, Principal Component Neural Networks, Wiley, New York, 1996.
[17] Y. LeCun, The MNIST database of handwritten digits, ⟨http://yann.lecun.com/exdb/mnist/⟩, 1998.
[18] C. Blake, C. Merz, UCI repository of machine learning databases, ⟨http://www.ics.uci.edu/∼mlearn/MLRepository.html⟩, 1998.
[19] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. USA 87 (1990) 9193–9196.
[20] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ⟨http://www.csie.ntu.edu.tw/∼cjlin/libsvm⟩, 2001.
[21] T. Takahashi, T. Kurita, Robust de-noising by kernel PCA, in: Proceedings of the International Conference on Artificial Neural Networks, Lecture Notes in Computer Science, vol. 2415, Springer, Berlin, 2002, pp. 739–744.
[22] C.-D. Lu, T.-Y. Zhang, X.-Z. Du, C.-P. Li, A robust kernel PCA algorithm, in: Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 5, IEEE, 2004, pp. 3084–3087.
[23] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, UK, 1993.
[24] C.J.C. Burges, Simplified support vector decision rules, in: L. Saitta (Ed.), Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, 1996, pp. 71–77.
[25] B. Schölkopf, P. Knirsch, A.J. Smola, C. Burges, Fast approximation of support vector kernel expansions, and an interpretation of clustering as approximation in feature spaces, in: P. Levi, R.-J. Ahlers, F. May, M. Schanz (Eds.), 20. DAGM Symposium Mustererkennung, Springer, Berlin, 1998, pp. 124–132.
[26] S. Mika, B. Schölkopf, A.J. Smola, K.-R. Müller, M. Scholz, G. Rätsch, Kernel PCA and de-noising in feature spaces, Adv. Neural Inf. Process. Syst. 11 (1999) 536–542.
[27] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Comput. 11 (1999) 443–482.
[28] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

About the Author—HEIKO HOFFMANN received his M.Sc. (Diploma) degree in physics from the University of Heidelberg in 2000 and his Ph.D. degree in computer science from the University of Bielefeld in 2004 while working at the Max Planck Institute for Human Cognitive and Brain Sciences in Munich. Currently, he is a research fellow at the Institute of Perception, Action and Behavior of the School of Informatics at Edinburgh University. His research interests include kernel methods, Bayesian learning, motor control in humanoids and humans, visual and tactile perception, and sensorimotor anticipation.