Classification with Kernel Mahalanobis Distance Classifiers

Bernard Haasdonk[1] and Elżbieta Pękalska[2]

[1] Institute of Numerical and Applied Mathematics, University of Münster, Germany, haasdonk@math.uni-muenster.de
[2] School of Computer Science, University of Manchester, United Kingdom, pekalska@cs.man.ac.uk

Abstract. Within the framework of kernel methods, linear data methods have almost completely been extended to their nonlinear counterparts. In this paper, we focus on nonlinear kernel techniques based on the Mahalanobis distance. Two approaches are distinguished here. The first one assumes an invertible covariance operator, while the second one uses a regularized covariance. We discuss conceptual and experimental differences between these two techniques and investigate their use in classification scenarios. For this, we involve a recent kernel method, called Kernel Quadratic Discriminant and, in addition, linear and quadratic discriminants in the dissimilarity space built by the kernel Mahalanobis distances. Experiments demonstrate the applicability of the resulting classifiers. The theoretical considerations and experimental evidence suggest that the kernel Mahalanobis distance derived from the regularized covariance operator is favorable.

Key words: Kernel Methods, Mahalanobis Distance, Quadratic Discriminant

1 Introduction

Nonlinear learning methods can be successfully designed by linear techniques in feature space induced by kernel functions. Many of such kernel methods have been proposed so far, including Support Vector Machine (SVM) and Kernel Fisher Discriminant (KFD) [2]. They have been widely applied to various learning scenarios thanks to their flexibility and good performance [6,7]. In this paper, we consider a nonlinear kernel technique, the kernel Mahalanobis distance, which represents a kernel quadratic analysis tool. Two approaches to kernel Mahalanobis distance are distinguished and investigated here. The first one assumes invertible class covariance matrices in the kernel-induced feature space and is similar to the method discussed in [5], while the other one regularizes them appropriately. As a result, these different assumptions lead to different formulations of kernel Mahalanobis classifiers. The goal of the current presentation is to compare these two approaches theoretically and experimentally. For the experiments we use different classifiers built on these

kernel Mahalanobis distances. First, we use Kernel Quadratic Discriminant (KQD) analysis [4]. We also train classifiers in simple dissimilarity spaces [3] defined by the class-wise kernel Mahalanobis distances. In this way, we make an explicit use of the between-class information, which may also lead to favorable results. Our approach KQD is a pure kernelized algorithm and differs from the two-stage approach [8] which relies on supervised dimension reduction in a kernel-induced space followed by a quadratic discriminant analysis.

The paper is organized as follows. Section 2 starts with preliminaries on kernels. Section 3 introduces the kernel Mahalanobis distances and subsequent classification strategies. Section 4 presents an experimental study on the kernel Mahalanobis distance classifiers on toy and real world data. Section 5 gives some theoretical insights and we conclude with Section 6.

2 Kernels and Feature-Space Embedding

Let $\mathcal{X}$ be a set of objects, either a vector space or a general set of structured objects. Let $\phi\colon \mathcal{X} \to \mathcal{H}$ be a mapping of patterns from $\mathcal{X}$ to a high-dimensional or infinite dimensional Hilbert space $\mathcal{H}$ with the inner product $\langle\cdot,\cdot\rangle$.

We address a c-class problem, given by the training data $X := \{x_i\}_{i=1}^n \subset \mathcal{X}$ with labels $\{y_i\}_{i=1}^n \subset \Omega$, where $\Omega := \{\omega_1, \ldots, \omega_c\}$ is a set of c target classes. Let $\Phi := [\phi(x_1), \ldots, \phi(x_n)]$ be the sequence of images of the training data X in $\mathcal{H}$. Given the embedded training data, the empirical mean is defined as $\phi_\mu := \frac{1}{n}\sum_{i=1}^n \phi(x_i) = \frac{1}{n}\Phi\mathbf{1}_n$, where $\mathbf{1}_n$ is an n-element vector of all ones. Here and in the following we will use such matrix-vector-product notation involving Φ for both finite and infinite dimensional $\mathcal{H}$ which is reasonable by suitable interpretation as linear combinations in $\mathcal{H}$. The mapped training data vectors are centered by subtracting their mean such that $\tilde{\phi}(x_i) := \phi(x_i) - \phi_\mu$, or, more compactly, $\tilde{\Phi} := [\tilde{\phi}(x_1), \ldots, \tilde{\phi}(x_n)] = \Phi - \frac{1}{n}\phi_\mu\mathbf{1}_n^\mathsf{T} = \Phi - \frac{1}{n}\Phi\mathbf{1}_n\mathbf{1}_n^\mathsf{T} = \Phi H$. Here, $H := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ is the $n \times n$ centering matrix, while I_n is the $n \times n$ identity matrix. Note that $H = H^\mathsf{T} = H^2$. The empirical covariance operator $C\colon \mathcal{H} \to \mathcal{H}$ acts on $\phi(x) \in \mathcal{H}$ as $C\,\phi(x) := \frac{1}{n}\sum_{i=1}^n (\phi(x_i) - \phi_\mu)\,\langle\phi(x_i) - \phi_\mu, \phi(x)\rangle = \frac{1}{n}\sum_{i=1}^n \tilde{\phi}(x_i)(\tilde{\phi}(x_i))^\mathsf{T}\phi(x) = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T}\phi(x)$. Here, we use the transpose notation $\phi(x)^\mathsf{T}\phi(x') := \langle\phi(x), \phi(x')\rangle$ as an abbreviation for inner products, hence $\tilde{\Phi}^\mathsf{T}\phi(x)$ denotes a column-vector of inner products. We can therefore interpret $\frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T}$ as an operator and identify the empirical covariance as $C = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T} = \frac{1}{n}\Phi HH\Phi^\mathsf{T}$.

The transformation ϕ acts as a (usually) nonlinear map to a high-dimensional space $\mathcal{H}$ in which the classification task can be handled in either a more efficient or more beneficial way. In practice, we will not necessarily know ϕ, but choose a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that encodes the inner product in $\mathcal{H}$, instead. The kernel k is a positive definite function such that $k(x, x') = \phi(x)^\mathsf{T}\phi(x')$ for any $x, x' \in \mathcal{X}$. Particular instances of such kernels are the Gaussian Radial Basis Function $k_{\mathrm{rbf}}(x, x') := \exp(-\gamma||x - x'||^2)$ for $\gamma \in \mathbb{R}_+$ and the polynomial kernel $k_{\mathrm{pol}} := (1 + \langle x, x'\rangle)^p$ for $p \in \mathbb{N}$. Given that $\mathcal{X} = \mathbb{R}^d$,

the kernel k_{rbf} represents an inner product in an infinite dimensional Hilbert space $\mathcal{H}$, in contrast to a finite dimensional space for the polynomial kernel k_{pol}. For details on kernel methods we refer to [6,7]. $K := \Phi^{\mathsf{T}}\Phi$ is an $n \times n$ kernel matrix derived from the training data. Moreover, we will also use the centered kernel matrix $\tilde{K} := \tilde{\Phi}^{\mathsf{T}}\tilde{\Phi} = H\Phi^{\mathsf{T}}\Phi H = HKH$. Further, for an arbitrary $x \in \mathcal{X}$, $\mathbf{k}_x := [k(x_1, x),\ \ldots,\ k(x_n, x)]^{\mathsf{T}} = \Phi^{\mathsf{T}}\phi(x)$ denotes the vector of kernel values of x to the training data, while $\tilde{\mathbf{k}}_x := \tilde{\Phi}^{\mathsf{T}}\tilde{\phi}(x) = H(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n)$ is the centered vector. Finally, we will also use the self-similarity $k_{xx} := k(x, x) = \phi(x)^{\mathsf{T}}\phi(x)$ and its centered version $\tilde{k}_{xx} = \tilde{\phi}(x)^{\mathsf{T}}\tilde{\phi}(x) = k_{xx} - \frac{2}{n}\mathbf{1}_n^{\mathsf{T}}\mathbf{k}_x + \frac{1}{n^2}\mathbf{1}_n^{\mathsf{T}}K\mathbf{1}_n$. In addition to the quantities defined for the complete sequence Φ, we can define analogous class-wise quantities which are indicated with the superscript $[j]$.

3 Kernel Mahalanobis Distance Classifiers

With the above notation, the Mahalanobis distance in the kernel-induced feature space $\mathcal{H}$ can be formulated purely in terms of kernel evaluations as we derive in the following. Then we introduce the subsequent classifiers.

3.1 Kernel Mahalanobis Distances for Invertible Covariance

For simplicity of presentation, we consider here a single class of n elements $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$. For classification, the resulting formulae will be used in a class-wise manner. We require here an invertible empirical class covariance operator C in the kernel-induced space. This limits our reasoning to a finite-dimensional $\mathcal{H}$, as the image of C based on n samples has a finite dimension $m < n$. We want to kernelize the empirical square Mahalanobis distance

$$d^2(\phi(x); \{\phi_\mu, C\}) := (\phi(x) - \phi_\mu)^{\mathsf{T}} C^{-1} (\phi(x) - \phi_\mu). \tag{1}$$

Since $\mathcal{H}$ is m-dimensional, with $m < n$, we may interpret $\tilde{\Phi}$ as an $m \times n$ matrix. Hence, it has a singular value decomposition $\tilde{\Phi} = USV^{\mathsf{T}}$ with orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $S \in \mathbb{R}^{m\times n}$. By using the orthogonality of U and V, we have: $C = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^{\mathsf{T}} = \frac{1}{n}USS^{\mathsf{T}}U^{\mathsf{T}}$ and $\tilde{K} = \tilde{\Phi}^{\mathsf{T}}\tilde{\Phi} = VS^{\mathsf{T}}SV^{\mathsf{T}}$, with an invertible matrix $SS^{\mathsf{T}} \in \mathbb{R}^{m\times m}$ but singular $S^{\mathsf{T}}S \in \mathbb{R}^{n\times n}$. So $C^{-1} = nU(SS^{\mathsf{T}})^{-1}U^{\mathsf{T}}$ and $\tilde{K}^{-} = V(S^{\mathsf{T}}S)^{-}V^{\mathsf{T}}$, where the superscript $^{-}$ denotes the pseudo-inverse. Multiplication of these equations with $\tilde{\Phi}$ yields $\frac{1}{n}C^{-1}\tilde{\Phi} = U(SS^{\mathsf{T}})^{-1}SV^{\mathsf{T}}$ and $\tilde{\Phi}\tilde{K}^{-} = US(S^{\mathsf{T}}S)^{-}V^{\mathsf{T}}$. Since $S \in \mathbb{R}^{m\times n}$ is diagonal and has m nonzero singular values, both middle matrices $(SS^{\mathsf{T}})^{-1}S$ and $S(S^{\mathsf{T}}S)^{-}$ are $m \times n$ diagonal matrices with inverted singular values on the diagonal. Therefore, these matrices are identical and we conclude that

$$\tilde{\Phi}\tilde{K}^{-} = \frac{1}{n}C^{-1}\tilde{\Phi}. \tag{2}$$

Given a centered vector $\tilde{\phi}(x) = \phi(x) - \frac{1}{n}\Phi\mathbf{1}_n$, C acts on $\tilde{\phi}(x)$ as follows:

$$C\tilde{\phi}(x) = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^{\mathsf{T}}\left(\phi(x) - \frac{1}{n}\Phi\mathbf{1}_n\right) = \frac{1}{n}\tilde{\Phi}H\left(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n\right) = \frac{1}{n}\tilde{\Phi}\tilde{\mathbf{k}}_x. \tag{3}$$

Since C is invertible, this implies with (2) that $\tilde{\phi}(x) = \frac{1}{n} C^{-1} \tilde{\Phi} \tilde{\mathbf{k}}_x = \tilde{\Phi} \tilde{K}^{-} \tilde{\mathbf{k}}_x$. Together with the identity (2) this allows us to express the Mahalanobis distance for invertible covariance operator in its kernelized form as:

$$d^2_{IC}(x) := d^2_{IC}(\phi(x); \{\phi_\mu, C\}) = \tilde{\phi}(x)^\mathsf{T} C^{-1} \tilde{\phi}(x) = n\, \tilde{\mathbf{k}}_x^\mathsf{T} (\tilde{K}^{-})^2 \tilde{\mathbf{k}}_x. \quad (4)$$

In practice the computation of $\tilde{K}^{-}$ relies on a threshold $\alpha > 0$ such that singular values smaller than α are treated as 0. Hence, the distance d^2_{IC} has a regularization parameter α, which must be chosen properly during training.

3.2 Kernel Mahalanobis Distance for Regularized Covariance

The empirical covariance operator may not be invertible as we work with finite samples in a high-dimensional/infinite dimensional space $\mathcal{H}$. As an ansatz we directly regularize the covariance operator to prevent it from being singular: $C_{\text{reg}} := C + \sigma^2 I_{\mathcal{H}} = \frac{1}{n} \tilde{\Phi} \tilde{\Phi}^\mathsf{T} + \sigma^2 I_{\mathcal{H}}$, where $\sigma^2 > 0$ is a parameter to be chosen. After multiplying by $\tilde{\Phi}$ from both sides, using $\tilde{K} = \tilde{\Phi}^\mathsf{T} \tilde{\Phi}$ and defining $\tilde{K}_{\text{reg}} := \tilde{K} + \alpha I_n$ for $\alpha := n\sigma^2$, we get $C_{\text{reg}} \tilde{\Phi} = \frac{1}{n} \tilde{\Phi} (\tilde{K} + n\sigma^2 I_n) = \frac{1}{n} \tilde{\Phi} \tilde{K}_{\text{reg}}$. As a result, both C_{reg} and $\tilde{K}_{\text{reg}}$ are strictly positive definite, hence non-singular, as $n\sigma^2 > 0$. The inverses are therefore well-defined, leading to an equivalent of (2) as

$$\tilde{\Phi} \tilde{K}^{-1}_{\text{reg}} = \frac{1}{n} C^{-1}_{\text{reg}} \tilde{\Phi}. \quad (5)$$

Note that C_{reg} acts on an arbitrary centered vector $\tilde{\phi}(x)$ as $C_{\text{reg}}\, \tilde{\phi}(x) = \frac{1}{n} \tilde{\Phi} \tilde{\mathbf{k}}_x + \sigma^2 \tilde{\phi}(x)$, directly following from (3). Since C_{reg} is invertible, we obtain

$$\tilde{\phi}(x) = \frac{1}{n} C^{-1}_{\text{reg}} \tilde{\Phi} \tilde{\mathbf{k}}_x + \sigma^2 C^{-1}_{\text{reg}} \tilde{\phi}(x). \quad (6)$$

After multiplying (6) on both sides by $\tilde{\phi}(x)^\mathsf{T}$ (from the left) and thanks to (5), we can write $\tilde{\phi}(x)^\mathsf{T} \tilde{\phi}(x) = \tilde{\phi}(x)^\mathsf{T} \tilde{\Phi} \tilde{K}^{-1}_{\text{reg}} \tilde{\mathbf{k}}_x + \sigma^2 \tilde{\phi}(x)^\mathsf{T} C^{-1}_{\text{reg}} \tilde{\phi}(x)$. We can solve for the desired square Mahalanobis distance in the last term. By using the kernel quantities $\tilde{k}_{xx} = \tilde{\phi}(x)^\mathsf{T} \tilde{\phi}(x)$ and $\tilde{\mathbf{k}}_x = \tilde{\Phi}^\mathsf{T} \tilde{\phi}(x)$ we obtain the kernel Mahalanobis distance for regularized covariance

$$d^2_{RC}(x) := d^2(\phi(x); \{\phi_\mu, C_{\text{reg}}\}) = \tilde{\phi}(x)^\mathsf{T} C^{-1}_{\text{reg}} \tilde{\phi}(x) = \frac{1}{\sigma^2} (\tilde{k}_{xx} - \tilde{\mathbf{k}}_x^\mathsf{T} K^{-1}_{\text{reg}} \tilde{\mathbf{k}}_x). \quad (7)$$

3.3 Classifiers Based on Kernel Mahalanobis Distances

Kernel Quadratic Discriminant (KQD). First, we consider the straightforward extension of Quadratic Discriminant (QD) analysis in Euclidean spaces. This leads to Kernel Quadratic Discriminants (KQD) [4]. For a c-class problem in a space $\mathcal{X} = \mathbb{R}^d$ with regular class-wise covariance matrices $\Sigma^{[j]}$, means $\mu^{[j]}$ and prior probabilities $P(\omega_j)$, the quadratic discriminant for the j-th class is given as $f^{[j]}(x) := -\frac{1}{2}(x - \mu^{[j]})^\mathsf{T} (\Sigma^{[j]})^{-1} (x - \mu^{[j]}) + b_j$, where

$b_j := -\frac{1}{2}\ln(\det(\Sigma^{[j]})) + \ln(P(\omega_j))$. A new sample x is classified to ω_i with $i = \arg\max_{j=1,\ldots,c} f^{[j]}(x)$; see for instance [1].

By inserting the class-wise kernel Mahalanobis distances, two different decision functions are obtained for KQD, $f^{[j]}_{IC}(x) := -(d^{[j]}_{IC}(x))^2 + b_j$ and $f^{[j]}_{RC}(x) := -(d^{[j]}_{RC}(x))^2 + b_j$ for the invertible and regularized covariance case, respectively. The offset b_j can be expressed by kernel evaluations thanks to $\ln(\det(C^{[j]}) = \ln\prod(\lambda^{[j]}_i)$ where the eigenvalues $\lambda^{[j]}_i$ of $C^{[j]}$ are identical to the eigenvalues of $\frac{1}{n_j}\tilde{K}^{[j]}$ for $i = 1, \ldots, l := \mathrm{rank}(\tilde{K}^{[j]})$. Numerical problems however arise in computing the logarithm of the eigenvalue-product, if many small eigenvalues occur. This happens in practice because a kernel matrix has often a slowly decaying eigenvalue spectrum. Consequently, we choose the offset values by a training error minimization procedure; see [4] for details. In the following we refer to the resulting classifiers as KQD-IC and KQD-RC.

Fisher and Quadratic Discriminants in Dissimilarity Spaces. We can define new features of a low-dimensional space by the square kernel Mahalanobis distances computed to the class means. Hence, given a c-class problem and class-wise squared dissimilarities $(d^{[j]}(x))^2, j = 1, \ldots, c$, we can define a data-dependent mapping to a c-dimensional dissimilarity space $\psi : \mathcal{X} \to \mathbb{R}^c$ with $\psi(x) := [(d^{[1]}(x))^2, \ldots (d^{[c]}(x))^2]^\mathsf{T}$. This can be done for either the d^2_{IC} or d^2_{RC} distances. For $c = 2$ classes, the KQD decision boundary is simply a line parallel to the main diagonal in this 2D dissimilarity space. For certain data distributions, more complex decision boundaries may be required. Since kernel Mahalanobis distances are derived based on the within-class information only, subsequent decision functions in this dissimilarity space enable us to use the between-class information more efficiently. Two classifiers are here considered, namely Fisher Discriminants (FD) and Quadratic Discriminant (QD); see e.g. [1]. Since we apply these in two dissimilarity spaces defined by either d^2_{IC} or d^2_{RC}, we get four additional classification strategies denoted as FD-IC, FD-RC, QD-IC and QD-RC, correspondingly.

4 Experiments

In order to get insights into the kernel Mahalanobis distances, we first perform 2D experiments on an artificial data set for different sample sizes and kernels. Then we target at some real-world problems. We include three reference classifiers to compare the overall classification performance. These are two linear kernel classifiers, Support Vector Machine (SVM) [6] and Kernel Fisher Discriminant (KFD) [2], and a nonlinear Kernel k-Nearest Neighbor (KNN) classifier. The KNN classifier is based on the kernel-induced distance in the feature space $||\phi(x) - \phi(x')||^2 = k(x,x) - 2k(x,x') + k(x',x')$, which corresponds to the usual k-nearest neighbor decision in the input space for a Gaussian kernel. The regularization parameters are the usual C for penalization in SVM, β for regularizing the within-class scatter in KFD and the number of neighbors k for KNN. All experiments rely on PRtools41 (`http://prtools.org`).

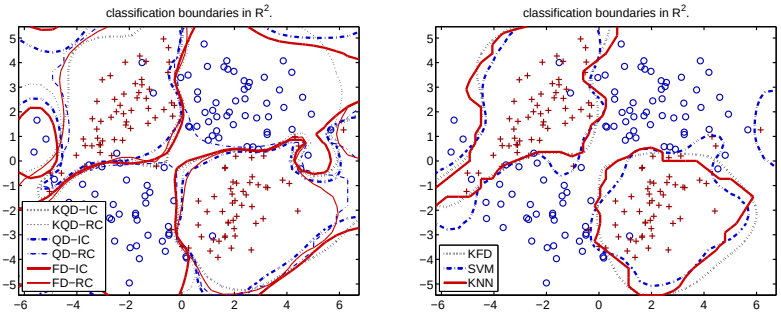

Fig. 1. Cross-validated classifiers on 2D toy data with kernel k_{rbf}.

4.1 Experiments on 2D Toy Data

We consider a two-class toy problem as illustrated in Fig. 1. Both classes have equal class-priors and are generated by a mixture of two normal distributions such that the resulting distributions are no longer unimodal. Hence, QD analysis is invalid here and stronger nonlinear models must be applied. The training set consists of 200 samples. We study both Gaussian and polynomial kernels, k_{rbf} and k_{pol}. The optimal kernel parameters and regularization parameters of the classifiers are chosen by 10-fold cross-validation. The cross validation range for the kernel parameters are $\gamma \in [0.01, 50]$ discretized by 8 values and $p = 1, 2, 3, 4$. The regularization parameters and cross-validation ranges (each discretized by 8 values) are $\alpha \in [10^{-6}, 10^{-1}]$ (class-wise identical) for KQD-IC, FD-IC and QD-IC, $\alpha = n\sigma^2 \in [10^{-5}, 1]$ (class-wise identical) for KQD-RC, FD-RC and QD-RC, $C \in [10^{-1}, 10^{6}]$ for SVM, $k \in [1, 8]$ for KNN and $\beta \in [10^{-6}, 10]$ for KFD. The resulting kernel Mahalanobis classifiers with kernel k_{rbf} and the training data set are depicted in the left plot of Fig. 1. The right plot shows the reference classifiers. The KNN rule is, as expected, highly nonlinear. Overall, all classifiers perform reasonably well.

The classification errors are determined on an independently drawn test set of 1000 examples. The procedure of data drawing, cross-validated training of the classifiers and test-error determination is repeated for ten random training and test-set drawings. The mean errors and standard deviations are shown in Table 1. To assess the dependence on the sample number, we also determine results for smaller training sample sizes n.

Among the reference classifiers we see that nonseparability is problematic for SVM as it performs worse than the KNN approach for pronounced cases (larger n). KFD is frequently similar or better than SVM, as also reported in other studies [2]. Among the different Mahalanobis distances we observe a superiority of the approaches based on d^2_{RC} over those using d^2_{IC}. The difference in performance is increasing with the decrease of the sample size n. KQD-IC seems favorable among the IC-approaches. Concerning the RC-approaches, QD-RC seems favorable for k_{rbf}, while KQD-RC seems favorable for k_{pol}. Good results are obtained by both k_{rbf} and k_{pol} for this data set. Compar-

Table 1. Average classification errors [in %] for 2D data with different training sample sizes n and kernels. Numbers in parenthesis denote standard deviations.

	k_{rbf}			k_{pol}		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
KQD-IC	20.8 (4.2)	17.4 (1.1)	15.5 (1.4)	18.8 (1.8)	17.2 (1.6)	16.0 (1.5)
FD-IC	20.9 (3.8)	17.9 (2.3)	15.7 (2.0)	20.8 (5.0)	19.3 (2.6)	16.0 (1.2)
QD-IC	21.7 (4.8)	16.7 (0.9)	16.0 (1.7)	19.7 (3.5)	18.4 (2.3)	17.3 (1.8)
KQD-RC	18.8 (2.1)	16.2 (1.0)	15.3 (1.6)	16.2 (1.8)	16.5 (1.8)	14.9 (1.2)
FD-RC	18.4 (2.1)	17.5 (1.8)	15.3 (1.7)	17.5 (2.9)	17.9 (2.7)	15.5 (1.0)
QD-RC	18.5 (2.2)	15.8 (1.2)	14.9 (1.8)	19.5 (4.2)	18.5 (3.1)	17.2 (1.9)
KFD	19.5 (3.1)	16.5 (2.2)	14.7 (1.4)	16.7 (2.3)	16.4 (2.4)	14.5 (1.2)
SVM	19.0 (2.0)	17.0 (1.8)	16.1 (2.7)	17.4 (2.4)	19.6 (6.3)	17.9 (2.2)
KNN	18.6 (3.0)	16.3 (1.6)	15.4 (1.6)	17.7 (2.8)	17.0 (2.5)	16.7 (1.4)

ing the kernel Mahalanobis approaches to the reference methods, the former provide similar results to those of the reference classifiers.

4.2 Real-World-Experiments

We use data from the UCI Repository (`http://archive.ics.uci.edu/ml/`). They describe problems with categorical, continuous and mixed features and with varying number of dimensions and classes. Each data set is split into training and test sets in the ratio of r_{tr} as specified in Table 2. We standardize the vectorial data and apply a Gaussian kernel k_{rbf}. For multiclass problems, SVM and KFD are trained in the one-vs-all scenario. As before, the optimal kernel parameter γ and regularization parameters of all classifiers are determined by 10-fold cross-validation with partially slightly adjusted search ranges, i.e. $\alpha \in [10^{-6}, 5 \cdot 10^{-1}]$ for KQD-IC, FLD-IC and QD-IC, $\alpha = n\sigma^2 \in [10^{-5}, 2]$ for KQD-RC, FLD-RC and QD-RC, $C \in [10^{-1}, 10^{6}]$ for SVM, $k \in [1, 15]$ for KNN and $\beta \in [10^{-6}, 2]$ for KFD. The average test-errors and the standard deviations over ten repetitions are reported in Table 3.

Concerning the reference methods, we observe that KFD is mostly best, sometimes outperformed by SVM. Among the kernel Mahalanobis classifiers we again note that the RC-versions are almost uniformly better than the IC-versions. In a number of cases the IC-versions are clearly inferior (Ecoli, Glass, Heart, Mfeat-*, Sonar, Wine, Ionosphere). This occurs when the number of samples is low as compared to the original dimensionality. Interestingly, QD-RC often gives similar or better results than KQD-RC, which is not analogous for the IC-versions. The kernel Mahalanobis classifiers are mostly comparable to the reference classifiers for both binary and multiclass problems. QD-RC performs overall the best (also better than reference classifiers) for the Diabetes, Imox, Ionosphere, and Wine data. Both KQD-RC and QD-RC classifiers are better than the reference classifiers for the Imox and Sonar data.

Table 2. Data used in our experiments and hold-out ratio r_{tr}.

Data	#Obj.	#Feat.	#Class	Class sizes	r_{tr}	Variables
Biomed	194	5	2	127/67	0.50	Mixed
Diabetes	768	8	2	500/268	0.50	Mixed
Ecoli	272	6	3	143/77/52	0.50	Continuous
Glass	214	9	4	70/76/17/51	0.50	Continuous
Heart	297	13	2	160/137	0.50	Mixed
Imox	192	8	4	48	0.50	Integer-valued
Ionosphere	351	34	2	225/126	0.50	Continuous
Liver	345	6	2	145/200	0.50	Cont. Integer-valued
Mfeat-Fac	2000	216	10	200	0.15	Continuous
Mfeat-Fou	2000	76	10	200	0.15	Continuous
Sonar	208	60	2	97/111	0.50	Continuous
Wine	178	13	3	59/71/48	0.50	Continuous

Table 3. Average classification errors [in %] for real data and kernel k_{rbf}. Numbers in parenthesis denote the standard deviations.

	Biomed	Diabetes	Ecoli	Glass	Heart	Imox
KQD-IC	16.2 (3.8)	28.3 (1.8)	7.6 (3.6)	46.7 (8.2)	20.5 (2.0)	7.2 (2.4)
FD-IC	22.6 (5.0)	32.6 (2.4)	12.0 (3.0)	49.8 (3.5)	21.1 (2.0)	14.1 (4.6)
QD-IC	16.5 (4.1)	29.6 (2.2)	7.2 (2.8)	52.0 (4.2)	21.9 (1.9)	8.4 (2.0)
KQD-RC	16.6 (3.1)	28.2 (2.1)	5.9 (1.6)	44.0 (6.3)	16.7 (1.9)	9.2 (3.4)
FD-RC	16.4 (4.4)	28.2 (1.2)	5.8 (1.8)	44.4 (4.3)	17.1 (2.4)	10.9 (3.8)
QD-RC	15.5 (2.8)	25.8 (2.3)	5.6 (1.9)	40.7 (4.8)	18.3 (2.7)	6.6 (2.5)
KFD	16.5 (2.8)	26.3 (2.1)	5.2 (1.6)	36.7 (5.7)	18.4 (2.3)	9.4 (2.2)
SVM	15.2 (2.3)	28.9 (2.3)	5.2 (2.3)	39.3 (5.0)	16.4 (2.3)	10.1 (3.3)
KNN	20.6 (3.7)	30.8 (0.9)	7.4 (2.1)	43.9 (5.3)	17.3 (2.6)	9.6 (5.3)

	Ionosphere	Liver	Mfeat-Fac	Mfeat-Fou	Sonar	Wine
KQD-IC	11.2 (2.6)	35.6 (4.0)	10.0 (1.7)	61.4 (3.2)	29.5 (5.7)	5.1 (2.6)
FD-IC	12.2 (2.5)	41.8 (3.8)	13.5 (1.4)	55.7 (3.4)	31.7 (4.8)	6.5 (2.8)
QD-IC	11.7 (2.2)	42.1 (3.7)	12.4 (1.8)	35.5 (2.7)	35.5 (3.1)	7.4 (3.3)
KQD-RC	7.8 (3.3)	39.6 (4.6)	6.1 (0.6)	25.1 (1.4)	15.7 (3.2)	3.8 (1.4)
FD-RC	7.5 (2.0)	37.6 (2.9)	7.1 (0.8)	26.6 (1.4)	22.0 (4.0)	3.5 (1.4)
QD-RC	5.8 (1.7)	39.6 (3.4)	6.1 (1.0)	25.7 (1.1)	16.6 (2.5)	2.8 (1.7)
KFD	6.8 (2.2)	32.9 (2.6)	3.9 (0.6)	22.9 (0.9)	17.7 (3.3)	3.8 (1.9)
SVM	7.1 (1.4)	30.4 (3.1)	4.7 (0.6)	23.0 (1.0)	18.2 (5.3)	3.1 (1.8)
KNN	23.9 (14.7)	41.2 (3.7)	8.1 (6.2)	28.3 (1.6)	19.8 (3.9)	8.3 (7.1)

5 Discussion and Theoretical Considerations

We focus now on some theoretical aspects concerning the kernel Mahalanobis distances with respect to their usage.

Assumption on Invertible Covariance. The motivation behind the distance d_{IC}^2 requires that the covariance operator is invertible. As a theoretical consequence, the sound derivation is limited to a finite dimensional $\mathcal{H}$. This is violated e.g. for the Gaussian kernel k_{rbf}. Counterintuitive situations may

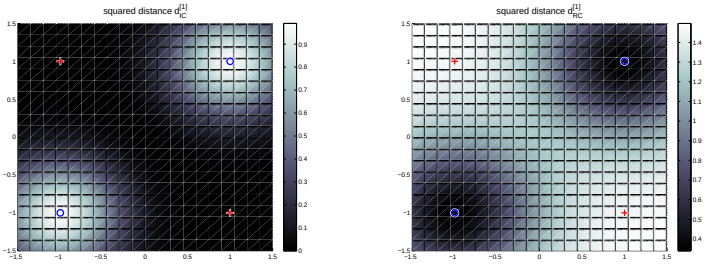

Fig. 2. XOR-example and square kernel Mahalanobis distances for the kernel k_{rbf}. The left plot shows $(d^{[1]}_{IC}(x))^2$, while the right plot shows $(d^{[1]}_{RC}(x))^2$.

occur if non-singularity does not hold: a vector $\tilde{\mathbf{k}}_x$ in (4) may be nonzero but lie in the eigenspace of $\tilde{K}$ corresponding to the eigenvalue 0. This may occur if x is atypical with respect to the training samples. Simple computation yields $d^2_{IC}(x) = 0$. If a classifier uses the distance as indication of a likelihood of x belonging to the corresponding class, the classification result will be clearly counterintuitive and possibly wrong. This phenomenon can be demonstrated on a simple 2-class XOR-data ($X = \{(-1,-1)^\mathsf{T}, (-1,1)^\mathsf{T}, (1,-1)^\mathsf{T}, (1,1)^\mathsf{T}\}, y = (\omega_1, \omega_2, \omega_2, \omega_1)$) as illustrated in Fig. 2, where the first class is plotted as circles and the second class as crosses. We plot a shading of the square kernel Mahalanobis distances of the first class resulting from the kernel k_{rbf} with $\gamma = 1$. We clearly see the fundamental qualitative difference between $d^{[1]}_{IC}$ (left, $\alpha = 10^{-4}$) and $d^{[1]}_{RC}$ (right, $\sigma^2 = 1$). The left plot demonstrates the problematic case, where the training examples of the first class have a higher distance to their own class than the samples of the second class. This illustrates why the discrimination power of the d^2_{IC} distances can decrease for few training samples in high-dimensional $\mathcal{H}$, as observed in our experiments. Nevertheless, the IC-methods are still applicable for infinite dimensional $\mathcal{H}$. Formally, the final decision rules are still well defined and can be applied independently whether the covariance operator is singular or not. Empirically, the results are frequently quite good. We may conclude that the pathological cases are rarely observed in practice if sufficiently many samples are available for training. Still a decrease in classification accuracy may be observed for few samples in high-dimensional spaces. In these cases, the use of d^2_{RC} is clearly more satisfactory and beneficial from a theoretical point of view.

Invariance. An interesting theoretical issue is invariance of the Mahalanobis distances in the kernel feature space. These invariance properties naturally transfer to kernel transformations that do not affect the resulting distances. One can easily check by definitions that the Mahalanobis distance is translation invariant in the feature space, i.e. $\bar{\phi}(x) := \phi(x) + \phi_0$ for a translation vector $\phi_0 \in \mathcal{H}$. Choosing $\phi_0 := \phi(x_0)$ for any $x_0 \in \mathcal{X}$ (or a general arbitrary linear combination) implies that both d^2_{IC} and d^2_{RC} remain identical by using the shifted kernel $\bar{k}(x,x') := \langle \bar{\phi}(x), \bar{\phi}(x') \rangle = k(x,x') + k(x,x_0) +$

$k(x', x_0) + k(x_0, x_0)$. In particular, kernel centering does not affect the distances. In analogy to Euclidean Mahalanobis distances, kernel Mahalanobis distances are invariant to scaling of the feature space by using the scaled kernels $\bar{k}(x, x') := \theta k(x, x')$ for $\theta > 0$. As we involve regularization parameters, this invariance only holds in practice if the regularization parameters are similarly scaled $\bar{\alpha} := \theta\alpha$ and $\bar{\sigma}^2 := \theta\sigma^2$. Consequently, a kernel can be used without a scale-parameter search.

6 Conclusion

We presented two versions of kernel Mahalanobis distance, d^2_{IC} and d^2_{RC}, derived either for invertible covariance operators or based on an additive regularization thereof. The distance d^2_{RC} leads to empirically better classification performance than d^2_{IC}, in particular for small sample size problems. Overall, the former measure is both conceptually and empirically favorable. These two Mahalanobis distances represent one-class models as only the within-class kernel information is used for their constructions. The between-class information is utilized in subsequent classifiers. Fully kernelized quadratic discriminant analysis can be performed by the KQD-IC/KQD-RC methods. Additional classifiers can be applied in the dissimilarity space obtained from the kernel Mahalanobis distances as illustrated with Fisher Discriminants FD-IC/FD-RC and Quadratic Discriminants QD-IC/QD-RC. Empirically, they often give comparable results to the reference classifiers. In several cases, QD-RC gives the overall best results. The kernel Mahalanobis classifiers can be advantageous for problems with high class overlap or nonlinear pattern distributions in a kernel-induced feature space.

References

1.R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001.
2.S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48, 1999.
3.E. Pękalska and R.P.W Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, 2005.
4.E. Pękalska and B. Haasdonk. Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. Accepted.
5.A. Ruiz and P.E. Lopez-de Teruel. Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 12(1):16–32, 2001.
6.B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.
7.J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, UK, 2004.
8.J. Wang, K.N. Plataniotis, J. Lu, and A.N. Venetsanopoulos. Kernel quadratic discriminant analysis for small sample size problem. *Pattern Recognition*, 41(5):1528–1538, 2008.