Kernel Discriminant Analysis for Positive Definite and Indefinite Kernels

Elżbieta Pekalska and Bernard Haasdonk

Abstract—Kernel methods are a class of well established and successful algorithms for pattern analysis due to their mathematical elegance and good performance. Numerous nonlinear extensions of pattern recognition techniques have been proposed so far based on the so-called kernel trick. The objective of this paper is twofold. First, we derive an additional kernel tool that is still missing, namely kernel quadratic discriminant (KQD). We discuss different formulations of KQD based on the regularized kernel Mahalanobis distance in both complete and class-related subspaces. Second, we propose suitable extensions of kernel linear and quadratic discriminants to indefinite kernels. We provide classifiers that are applicable to kernels defined by any symmetric similarity measure. This is important in practice because problem-suited proximity measures often violate the requirement of positive definiteness. As in the traditional case, KQD can be advantageous for data with unequal class spreads in the kernel-induced spaces, which cannot be well separated by a linear discriminant. We illustrate this on artificial and real data for both positive definite and indefinite kernels.

Index Terms—Machine learning, pattern recognition, kernel methods, indefinite kernels, discriminant analysis.

1 Introduction

Kernel methods are powerful statistical learning techniques [38], [36], widely applied to various learning scenarios due to their flexibility and good performance. A kernel is a (conditionally) positive definite (pd) function $k(x, x')$ of two variables x and x', and interpreted as a generalized inner product, hence natural similarity, in a reproducing kernel Hilbert space $\mathcal{H}$ induced by k [33], [40]. Due to the reproducing property of k, kernel-based classifiers are indirectly built in $\mathcal{H}$ and often expressed as linear combinations of kernel values. Many traditional learning methods have been proposed so far in their kernel-based formulations. These include Support Vector Machines (SVM), kernel PCA, kernel Fisher discriminant (KFD), kernel k-means, and so on [36]. An additional tool that is still missing within the set of simple approaches is the kernel quadratic discriminant (KQD). In this paper, we derive KQD as a natural extension of the quadratic discriminant in a Euclidean space. Three variants are considered in either full or class-related kernel-induced subspaces.

Although traditional kernel methods have now been applied to general nonvectorial data descriptions, such as strings, bags of words, graphs, shapes, probability models [35], [36], the class of permissible kernels is often, and frequently wrongly, considered to be limited due to their requirement of being positive definite. In practice, however, many non-pd similarity measures arise, e.g., when invariance or robustness is incorporated into the measure [37], [20], [13]. Further reasons may include suboptimal optimization procedures for measure derivation [28], partial projections or occlusions [20], and context-dependent alignments or object comparisons [6], [30]. Naturally, indefinite (dis)similarities arise from non-euclidean or nonmetric dissimilarities, such as modified Hausdorff distances [6], or non-pd similarities, such as Kullback-Leibler divergence between probability distributions. Consequently, there is a practical need to handle these measures properly. In the case of metric dissimilarity measures, these can be embedded in Banach spaces where learning algorithms such as large margin classifiers can be applied [39], [16], [4]. Although these techniques provide alternatives to certain kernel methods for metric data, more general approaches are needed.

While many researchers choose to regularize non-pd kernels to make them pd, a natural extension of Mercer kernels leads to indefinite or Krein kernels [2], [25], [21], [11], [26], or dyadic kernels [18]. Both are examples of proximity representations, i.e., matrices whose elements encode degrees of similarity between pairs of objects and optimized prototypes [26]. Therefore, it is of high interest to develop and investigate methods that work with indefinite kernels. And indeed, an additional contribution of this paper is a sound underpinning of the approaches which extend kernel linear and quadratic discriminants to deal with indefinite kernels. Experiments on toy and real-world data show good performance of the KQD methods for both positive definite and indefinite kernels.

The paper is organized as follows: Section 2 starts with preliminaries on kernels. Section 3 presents the indefinite kernel Fisher discriminant analysis. Section 4 is the main part that introduces different formulations of KQD analysis for both positive definite and indefinite kernels. Section 5 focuses on an experimental study illustrating the performance of kernel discriminant analysis on toy and real-world

- *E. Pekalska is with the School of Computer Science, University of Manchester, Oxford Road, M13 9PL Manchester, UK. E-mail: pekalska@cs.man.ac.uk.*
- *B. Haasdonk is with the Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany. E-mail: haasdonk@mathematik.uni-stuttgart.de.*

Manuscript received 10 July 2008; revised 13 Aug. 2008; accepted 14 Nov. 2008; published online 2 Dec. 2008.
Recommended for acceptance by L. Bottou.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-07-0411.
Digital Object Identifier no. 10.1109/TPAMI.2008.290.

data. The final discussion is presented in Section 6. Due to space restrictions and in order to maintain clarity, the detailed derivations of the methods are left out from the main text, but provided as supplementary material in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290.

2 Preliminaries on Kernels

We will first introduce some notation and provide basic facts on Hilbert spaces and positive definite kernels. We will then focus on Krein spaces and the related indefinite kernels.

2.1 Positive Definite Kernels

Assume that $\mathcal{X}$ is a collection of objects $\{x\}$, either an index set, a set of original objects, or their vector representations in some input space. 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 $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Here, we will use notation that extends matrix-vector multiplications to Hilbert spaces. For two functions $\xi_1, \xi_2 \in \mathcal{H}$, we will equivalently write $\xi_1^{\mathrm{T}}\xi_2 := \langle\xi_1,\xi_2\rangle_{\mathcal{H}}$. A sequence of m vectors in $\mathcal{H}$ is denoted by $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_m]$. Given a vector $\mathbf{v}\in\mathbb{R}^m$, we define $\boldsymbol{\xi}\mathbf{v} := \sum_{i=1}^m v_i\xi_i$ as an abbreviation of linear combinations. Similarly, for a matrix $V = [\mathbf{v}_1,\ldots,\mathbf{v}_n]\in\mathbb{R}^{m\times n}$, $\boldsymbol{\xi}V := [\boldsymbol{\xi}\mathbf{v}_1,\ldots,\boldsymbol{\xi}\mathbf{v}_n]$ is a sequence of linear combinations defined by the columns of V. Hence, $\xi\mathbf{v}^{\mathrm{T}} = [v_1\xi,\ldots,v_m\xi]$ for a single $\xi\in\mathcal{H}$ and a vector $\mathbf{v}$. For two sequences $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_m]$ and $\boldsymbol{\xi}' = [\xi'_1,\ldots,\xi'_n]$ in $\mathcal{H}$, we will write $G := \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}'\in\mathbb{R}^{m\times n}$ to denote a cross-Gram matrix with entries $G_{ij} = \langle\xi_i,\xi'_j\rangle_{\mathcal{H}}$.

In this paper we address a c-class problem, given by the training data $X_{\mathrm{tr}} = \{x_i\}_{i=1}^n \subset \mathcal{X}$ with the corresponding 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 a sequence of images of the training data X_{tr} in $\mathcal{H}$. Without loss of generality, we assume that the vectors in Φ are grouped into classes such that $\Phi = [\Phi^{[1]},\Phi^{[2]},\ldots,\Phi^{[c]}]$, where $\Phi^{[j]} := [\phi(x_1^j),\ldots,\phi(x_{n_j}^j)]$ represents the jth class ω_j with n_j elements, which implies $\sum_{j=1}^c n_j = n$.

Given the training data $\Phi = [\phi(x_1),\ldots,\phi(x_n)]$, 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. The mapped training data vectors are centered by subtracting their mean such that $\tilde{\phi}(x_i) := \phi(x_i) - \phi_\mu$, or equivalently, $\tilde{\Phi} := [\tilde{\phi}(x_1),\ldots,\tilde{\phi}(x_n)] = \Phi - \phi_\mu\mathbf{1}_n^{\mathrm{T}} = \Phi - \frac{1}{n}\Phi\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} = \Phi H$. Here, $H := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is the $n\times n$ centering matrix, while I_n is the $n\times n$ identity matrix. H is symmetric, $H = H^{\mathrm{T}}$, and idempotent, $H = H^2$. The empirical covariance operator $C\colon \mathcal{H}\to\mathcal{H}$ is a continuous linear map defined by its operation on $\phi(x)\in\mathcal{H}$ as $C\,\phi(x) := \frac{1}{n}\sum_{i=1}^n(\phi(x_i)-\boldsymbol{\phi}_\mu)\,\langle\phi(x_i)-\boldsymbol{\phi}_\mu, \phi(x)\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n \tilde{\phi}(x_i)(\tilde{\phi}(x_i))^{\mathrm{T}}\,\phi(x) = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^{\mathrm{T}}\phi(x)$. We can therefore interpret $\frac{1}{n}\tilde{\Phi}\tilde{\Phi}^{\mathrm{T}}$ as an operator and identify the empirical covariance C as

$$C = \frac{1}{n}\,\tilde{\Phi}\tilde{\Phi}^{\mathrm{T}} = \frac{1}{n}\Phi HH\Phi^{\mathrm{T}}.$$

Given that the empirical covariance operator is invertible and $D_M^2(\cdot\,;\{\phi_\mu,C\})\colon \mathcal{H}\to\mathbb{R}_{\geq 0}$ is the empirical square Mahalanobis distance defined for a vector $\phi(x)\in\mathcal{H}$ as

$$D_M^2(\phi(x);\{\phi_\mu,C\}) := (\phi(x)-\phi_\mu)^{\mathrm{T}}\,C^{-1}(\phi(x)-\phi_\mu). \quad (1)$$

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 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 (semi)definite function such that $k(x,x') = \phi(x)^{\mathrm{T}}\phi(x')$ for any $x,x'\in\mathcal{X}$. Consequently, $K := \Phi^{\mathrm{T}}\Phi$ is an $n\times n$ kernel matrix derived from the training data. Moreover, we will also use the centered kernel matrix $\tilde{K} := \tilde{\Phi}^{\mathrm{T}}\tilde{\Phi} = H\Phi^{\mathrm{T}}\Phi H = HKH$. In addition to the quantities defined for the complete training sequence Φ, we can define analogous classwise quantities for $\Phi^{[j]}$, $j = 1,\ldots,c$, which are consequently indicated with the superscript $[j]$. Further on, for an arbitrary $x\in\mathcal{X}$, $\mathbf{k}_x$ denotes an n-element vector of kernel values of x to the training data, while $\tilde{\mathbf{k}}_x$ is the centered vector:

$$\begin{aligned}\mathbf{k}_x &:= [k(x_1,x),\ldots,k(x_n,x)]^{\mathrm{T}} = \Phi^{\mathrm{T}}\phi(x),\\ \tilde{\mathbf{k}}_x &:= \tilde{\Phi}^{\mathrm{T}}\tilde{\phi}(x) = H\left(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n\right).\end{aligned} \quad (2)$$

Finally, we will also make use of the self-similarity k_{xx} and its centered version $\tilde{k}_{xx}$:

$$\begin{aligned}k_{xx} &:= k(x,x) = \phi(x)^{\mathrm{T}}\phi(x),\\ \tilde{k}_{xx} &:= \tilde{\phi}(x)^{\mathrm{T}}\tilde{\phi}(x) = k_{xx} - \frac{2}{n}\mathbf{1}_n^{\mathrm{T}}\mathbf{k}_x + \frac{1}{n^2}\mathbf{1}_n^{\mathrm{T}}K\mathbf{1}_n.\end{aligned} \quad (3)$$

2.2 Indefinite Kernels

The terminology and notation presented in Section 2.1 can be extended to Krein spaces (see [1], [5], [31] for details). Note that, apart from pattern recognition [9], [26], also other fields such as H^∞ control [15] make use of linear estimation in Krein spaces. A *Kreĭn space* over $\mathbb{R}$ is a vector space $\mathcal{K}$ equipped with an indefinite inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}\colon \mathcal{K}\times\mathcal{K}\to\mathbb{R}$ such that $\mathcal{K}$ admits an orthogonal decomposition as a direct sum $\mathcal{K} = \mathcal{K}_+\oplus\mathcal{K}_-$, where $(\mathcal{K}_+,\langle\cdot,\cdot\rangle_+)$ and $(\mathcal{K}_-,\langle\cdot,\cdot\rangle_-)$ are separable Hilbert spaces with their corresponding positive definite inner products. The inner product of $\mathcal{K}$, however, is the difference of $\langle\cdot,\cdot\rangle_+$ and $\langle\cdot,\cdot\rangle_-$, i.e., for any $\xi_+,\xi'_+\in\mathcal{K}_+$ and any $\xi_-,\xi'_-\in\mathcal{K}_-$ holds

$$\langle\xi_+ + \xi_-,\xi'_+ + \xi'_-\rangle_{\mathcal{K}} := \langle\xi_+,\xi'_+\rangle_+ - \langle\xi_-,\xi'_-\rangle_-.$$

The decomposition is orthogonal with respect to this inner product, i.e., $\langle\xi_+,\xi_-\rangle_{\mathcal{K}} = 0$ for any $\xi_+\in\mathcal{K}_+$ and $\xi_-\in\mathcal{K}_-$. In particular, $\langle\xi_+,\xi_+\rangle_{\mathcal{K}} > 0$ and $\langle\xi_-,\xi_-\rangle_{\mathcal{K}} < 0$ for any nonzero vectors $\xi_+\in\mathcal{K}_+$ and $\xi_-\in\mathcal{K}_-$. Therefore, $\mathcal{K}_+$ is a *positive subspace*, while $\mathcal{K}_-$ is a *negative subspace*.

The orthogonal projections P_+ onto $\mathcal{K}_+$ and P_- onto $\mathcal{K}_-$ are called *fundamental projections*. Any $\xi\in\mathcal{K}$ can be represented as $\xi = P_+\xi + P_-\xi$, while $I_{\mathcal{K}} = P_+ + P_-$ is the identity operator. The linear operator $\mathcal{J} = P_+ - P_-$ is called the *fundamental symmetry* and is the basic characteristic of a Krein space $\mathcal{K}$, satisfying $\mathcal{J} = \mathcal{J}^{-1} = \mathcal{J}^{\mathrm{T}}$. The space $\mathcal{K}$ can be turned into its *associated Hilbert space* $|\mathcal{K}|$ by using the positive definite inner product $\langle\xi,\xi'\rangle_{|\mathcal{K}|} := \langle\xi,\mathcal{J}\xi'\rangle_{\mathcal{K}}$. Countable orthonormal

bases $\mathcal{B}_+$ for $\mathcal{K}_+$ and $\mathcal{B}_-$ for $\mathcal{K}_-$ give rise to a basis $\mathcal{B} := \mathcal{B}_+ \cup \mathcal{B}_-$ for $\mathcal{K}$. The latter is orthonormal in the sense that $\langle e, e' \rangle_{\mathcal{K}} = 0$ for all $e \neq e' \in \mathcal{B}$, $\langle e, e \rangle_{\mathcal{K}} = 1$ for all $e \in \mathcal{B}_+$, and $\langle e, e \rangle_{\mathcal{K}} = -1$ for all $e \in \mathcal{B}_-$. Similarly, as in the positive definite case, we use the "transposition" abbreviation $\xi^{\mathrm{T}} \xi' := \langle \xi, \xi' \rangle_{|\mathcal{K}|}$, and now additionally (motivated by $\mathcal{J}$ operating as a sort of "conjugation"), a "conjugate-transposition" notation $\xi^* \xi' := \langle \xi, \xi' \rangle_{\mathcal{K}} = \langle \mathcal{J}\xi, \xi' \rangle_{|\mathcal{K}|} = (\mathcal{J}\xi)^{\mathrm{T}} \xi' = \xi^{\mathrm{T}} \mathcal{J} \xi'$.

Finite-dimensional Krein spaces with $\mathcal{K}_+ = \mathbb{R}^p$ and $\mathcal{K}^- = \mathbb{R}^q$ are denoted by $\mathbb{R}^{(p,q)}$ and called *pseudo-euclidean spaces*. They are characterized by the so-called *signature* $(p, q) \in \mathbb{N}^2$. $\mathcal{J}$ becomes the matrix $\mathcal{J} = \mathrm{diag}(\mathbf{1}_p, -\mathbf{1}_q)$ with respect to an orthonormal basis in $\mathbb{R}^{(p,q)}$. Krein spaces are important as they provide feature-space representations of dissimilarity data [9] or indefinite kernels. For indefinite kernels, i.e., symmetric functions $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and finite data $\mathcal{X}$, the resulting kernel matrix K yields an embedding $\psi : \mathcal{X} \to \mathcal{K}$ into a finite-dimensional Krein space by its eigenvalue decomposition, such that $k(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{K}}$. In analogy to the pd case, an indefinite kernel represents an inner product in an implicitly defined feature space. Hence, algorithms working with indefinite kernels have a geometric interpretation in these spaces.

Let $\psi \colon \mathcal{X} \to \mathcal{K}$ be a mapping of the data into a Krein space $\mathcal{K}$ and $\Psi := [\psi(x_1), \ldots, \psi(x_n)]$ be a sequence of images of X_{tr} in $\mathcal{K}$. In the following, we adopt the matrix-vector multiplication notation from the previous section. All quantities derived in Section 2.1 can now be defined analogously, i.e., $\{\phi, \Phi, \phi_\mu\}$ are replaced by $\{\psi, \Psi, \psi_\mu\}$, inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ are replaced by $\langle \cdot, \cdot \rangle_{\mathcal{K}}$, transpositions ξ^{T} are replaced by conjugate-transpositions ξ^*, but transpositions of vectors $\mathbf{v} \in \mathbb{R}^n$ are maintained. In particular, the empirical mean is defined as $\psi_\mu := \frac{1}{n} \sum_{i=1}^n \psi(x_i) = \frac{1}{n} \Psi \mathbf{1}_n$. The data vectors in $\mathcal{K}$ are centered such that $\tilde{\psi}(x_i) := \psi(x_i) - \psi_\mu$; hence, $\tilde{\Psi} := [\tilde{\psi}(x_1), \ldots, \tilde{\psi}(x_n)] = \Psi - \frac{1}{n} \Psi \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}} = \Psi H$. The empirical covariance operator $C \colon \mathcal{K} \to \mathcal{K}$ is a continuous linear map that acts on $\psi(x) \in \mathcal{K}$ as $C\psi(x) := \frac{1}{n} \sum_{i=1}^n (\psi(x_i) - \psi_\mu) \langle \psi(x_i) - \psi_\mu, \psi(x) \rangle_{\mathcal{K}} = \frac{1}{n} \sum_{i=1}^n \tilde{\psi}(x_i) \langle \tilde{\psi}(x_i), \psi(x) \rangle_{\mathcal{K}} = \frac{1}{n} \sum_{i=1}^n \tilde{\psi}(x_i) \tilde{\psi}(x_i)^* \psi(x) = \frac{1}{n} \tilde{\Psi} \tilde{\Psi}^* \psi(x)$. We will therefore identify the empirical covariance operator as

$$C = \frac{1}{n} \tilde{\Psi} \tilde{\Psi}^* = \frac{1}{n} \tilde{\Psi} \tilde{\Psi}^{\mathrm{T}} \mathcal{J} = C^{|\mathcal{K}|} \mathcal{J},$$

where $C^{|\mathcal{K}|} = \frac{1}{n} \tilde{\Psi} \tilde{\Psi}^{\mathrm{T}}$ is the empirical covariance operator in $|\mathcal{K}|$. The operator C is not positive definite in the Hilbert sense, but it is in the Krein sense [1], [31]. It means that $\langle \xi, C\xi \rangle_{\mathcal{K}} \geq 0$ for $\xi \neq 0$, hence in agreement with the inner product of that space. Assuming C is invertible (which requires that $n > \dim(\mathcal{K})$), the empirical square Mahalanobis distance $D_M^2(\cdot; \{\psi_\mu, C\}) \colon \mathcal{K} \to \mathbb{R}_{\geq 0}$ of a vector $\psi(x) \in \mathcal{K}$ to the data described by the model $\{\psi_\mu, C\}$ is defined as

$$D_M^2(\psi(x); \{\psi_\mu, C\}) := (\psi(x) - \psi_\mu)^* \, C^{-1} (\psi(x) - \psi_\mu).$$

Since K represents the kernel matrix with respect to the inner product in $\mathcal{K}$, we get $K := \Psi^* \Psi = \Psi^{\mathrm{T}} \mathcal{J} \Psi$. Similar to traditional kernels, the centered kernel matrix is $\tilde{K} := \tilde{\Psi}^* \tilde{\Psi} = \tilde{\Psi}^{\mathrm{T}} \mathcal{J} \tilde{\Psi} = HKH$. Analogously, definitions (2) and (3) of $\mathbf{k}_x$, $\tilde{\mathbf{k}}_x$, k_{xx}, and $\tilde{k}_{xx}$ can be extended to indefinite kernels by suitable replacements. Table 1 summarizes these definitions for both types of kernels. In particular, $\mathcal{J} = I_{\mathcal{K}}$ in the positive definite case; hence, $\xi^* = \xi^{\mathrm{T}}$ and all definitions presented here reduce to the ones from Section 2.1. Note that we could have focused on the mere indefinite notation as the pd case is just a special instance. This would, however, have hampered the reading of subsequent sections and the distinction between the positive definite and indefinite parts. Consequently, we deliberately use ψ and Ψ in the indefinite case in contrast to ϕ and Φ from the pd case to make this distinction more obvious.

TABLE 1
Kernel-Induced Quantities for Positive Definite and Indefinite Kernels

Positive definite	Indefinite
$\tilde{\Phi} = \Phi H$	$\tilde{\Psi} = \Psi H$
$\phi_\mu = \frac{1}{n} \Phi \mathbf{1}_n$	$\psi_\mu = \frac{1}{n} \Psi \mathbf{1}_n$
$C = \frac{1}{n} \tilde{\Phi} \tilde{\Phi}^{\mathrm{T}}$	$C = \frac{1}{n} \tilde{\Psi} \tilde{\Psi}^*$
$K = \Phi^{\mathrm{T}} \Phi$	$K = \Psi^* \Psi$
$\tilde{K} = \tilde{\Phi}^{\mathrm{T}} \tilde{\Phi}$	$\tilde{K} = \tilde{\Psi}^* \tilde{\Psi}$
$\mathbf{k}_x = \Phi^{\mathrm{T}} \phi(x)$	$\mathbf{k}_x = \Psi^* \psi(x)$
$\tilde{\mathbf{k}}_x = H(\mathbf{k}_x - \frac{1}{n} K \mathbf{1}_n)$	$\tilde{\mathbf{k}}_x = H(\mathbf{k}_x - \frac{1}{n} K \mathbf{1}_n)$
$k_{xx} = \phi(x)^{\mathrm{T}} \phi(x)$	$k_{xx} = \psi(x)^* \psi(x)$
$\tilde{k}_{xx} = k_{xx} - \frac{2}{n} \mathbf{1}_n^{\mathrm{T}} \mathbf{k}_x + \frac{1}{n^2} \mathbf{1}_n^{\mathrm{T}} K \mathbf{1}_n$	$\tilde{k}_{xx} = k_{xx} - \frac{2}{n} \mathbf{1}_n^{\mathrm{T}} \mathbf{k}_x + \frac{1}{n^2} \mathbf{1}_n^{\mathrm{T}} K \mathbf{1}_n$

Data embeddings Φ and Ψ refer to kernel-induced Hilbert and Krein spaces, respectively.

3 Kernel Fisher Discriminant Analysis

Kernel Fisher discriminant (KFD) was proposed and successfully applied by Mika et al. [23], [24]. Since it is well known and due to space limits, we will directly focus on the extension to the indefinite case.

3.1 Indefinite Kernel Fisher Discriminant

Assume the training data for a two-class problem, $c = 2$, is embedded into a Krein space $\mathcal{K}$ by the mapping ψ, i.e., $\Psi := [\psi(x_1), \ldots, \psi(x_n)]$ is the sequence of mapped training data and $\psi_\mu^{[1]}, \psi_\mu^{[2]} \in \mathcal{K}$ are the class means. The Fisher linear discriminant attempts to find a direction $w \in \mathcal{K}$ such that the between-class scatter is maximized while the within-class scatter is minimized along w. In analogy to the positive definite case, the indefinite Fisher linear discriminant

$$f(x) = \langle w, \psi(x) \rangle_{\mathcal{K}} + b = w^* \psi(x) + b \tag{4}$$

is defined by the vector w that maximizes the Fisher criterion

$$J(w) = \frac{\langle w, \Sigma_B^{\mathcal{K}} w \rangle_{\mathcal{K}}}{\langle w, \Sigma_W^{\mathcal{K}} w \rangle_{\mathcal{K}}} = \frac{w^* \Sigma_B^{\mathcal{K}} w}{w^* \Sigma_W^{\mathcal{K}} w}, \tag{5}$$

where the between-class scatter operator acts as $\Sigma_B^{\mathcal{K}} w = (\psi_\mu^{[1]} - \psi_\mu^{[2]}) \langle \psi_\mu^{[1]} - \psi_\mu^{[2]}, w \rangle_{\mathcal{K}} = (\psi_\mu^{[1]} - \psi_\mu^{[2]})(\psi_\mu^{[1]} - \psi_\mu^{[2]})^{\mathrm{T}} \mathcal{J} w$. Hence, $\Sigma_B^{\mathcal{K}} = \Sigma_B^{|\mathcal{K}|} \mathcal{J}$, where $\Sigma_B^{|\mathcal{K}|} = (\psi_\mu^{[1]} - \psi_\mu^{[2]})(\psi_\mu^{[1]} - \psi_\mu^{[2]})^{\mathrm{T}}$ is the Hilbert between-class scatter operator in $|\mathcal{K}|$. Similarly, the within-class scatter operator can be expressed as $\Sigma_W^{\mathcal{K}} := \Sigma_W^{|\mathcal{K}|} \mathcal{J}$ with the Hilbert within-class scatter operator $\Sigma_W^{|\mathcal{K}|} := \sum_{j=1}^2 P(\omega_j) \sum_i (\psi(x_i^j) - \psi_\mu^{[j]})(\psi(x_i^j) - \psi_\mu^{[j]})^{\mathrm{T}}$ based on suitable

estimates of prior probabilities $P(\omega_j)$. The bias in the classifier can be chosen as $b = -\frac{1}{2}\langle w, \psi_\mu^{[1]} + \psi_\mu^{[2]} \rangle_{\mathcal{K}} = -\frac{1}{2} w^T \mathcal{J}(\psi_\mu^{[1]} + \psi_\mu^{[2]})$, such that the midpoint of $\psi_\mu^{[1]}$ and $\psi_\mu^{[2]}$ is on the decision line. The Fisher criterion can therefore be rewritten as

$$J(w) = \frac{w^{\mathrm{T}} \mathcal{J} \Sigma_B^{|\mathcal{K}|} \mathcal{J} w}{w^{\mathrm{T}} \mathcal{J} \Sigma_W^{|\mathcal{K}|} \mathcal{J} w}. \tag{6}$$

An important insight at this point is a geometric interpretation of the indefinite Fisher discriminant: Inserting the operator representations and substituting $v := \mathcal{J} w$ into the Fisher criterion (6) and the discriminant function (4) yields $J(w) = v^{\mathrm{T}} \Sigma_B^{|\mathcal{K}|} v / (v^{\mathrm{T}} \Sigma_W^{|\mathcal{K}|} v)$ and $f(x) = v^{\mathrm{T}} \psi(x) + b$ with b defined as $b = -\frac{1}{2} v^{\mathrm{T}} (\psi_\mu^{[1]} + \psi_\mu^{[2]})$. This means that the Fisher discriminant in the Krein space $\mathcal{K}$ is identical to the Fisher discriminant in the associated Hilbert space $|\mathcal{K}|$. This is by far not clear a priori and not valid for other indefinite kernel classifiers, e.g., indefinite SVM [11].

A kernel method should avoid such explicit embeddings into a Krein space and constructions of new inner products based on eigendecompositions. The kernel function should be used instead. And indeed, the discriminant can be obtained in a kernelized form by using the original indefinite kernel. Assume that the indefinite kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ encodes the inner product $k(x_i, x_j) = \psi(x_i)^{\mathrm{T}} \mathcal{J} \psi(x_j)$ in $\mathcal{K}$. As a result, the kernel matrix for the training data is $K = \Psi^{\mathrm{T}} \mathcal{J} \Psi$. Since $\Psi = [\Psi^{[1]}, \Psi^{[2]}]$, we can decompose $K = [K_1, K_2]$, where K_j is an $n \times n_j$ kernel submatrix for the jth class. The normal w can be written as an expansion of the form $w = \sum_{i=1}^n \alpha_i \psi(x_i) = \Psi \boldsymbol{\alpha}$. As a result, the indefinite kernel Fisher discriminant (IKFD) can be expressed as $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x) + b$. Moreover, given that $\mathbf{z} := [\frac{1}{n_1} \mathbf{1}_{n_1}^{\mathrm{T}}, -\frac{1}{n_2} \mathbf{1}_{n_2}^{\mathrm{T}}]^{\mathrm{T}}$ is an $n \times 1$ vector and $M := (K\mathbf{z})(K\mathbf{z})^{\mathrm{T}}$, we have

$$\begin{aligned} w^{\mathrm{T}} \mathcal{J} \Sigma_B^{|\mathcal{K}|} \mathcal{J} w &= \boldsymbol{\alpha}^{\mathrm{T}} \Psi^{\mathrm{T}} \mathcal{J} (\Psi \mathbf{z}) (\mathbf{z}^{\mathrm{T}} \Psi^{\mathrm{T}}) \mathcal{J} \Psi \alpha \\ &= \boldsymbol{\alpha}^{\mathrm{T}} (K\mathbf{z})(K\mathbf{z})^{\mathrm{T}} \alpha = \boldsymbol{\alpha}^{\mathrm{T}} M \boldsymbol{\alpha}. \end{aligned}$$

Similarly, we can derive that $w^{\mathrm{T}} \mathcal{J} \Sigma_W^{|\mathcal{K}|} \mathcal{J} w = \boldsymbol{\alpha}^{\mathrm{T}} N \boldsymbol{\alpha}$, where $N := \sum_{j=1}^2 P(\omega_j) K_j H^{[j]} K_j^{\mathrm{T}}$ and $H^{[j]} = I_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^{\mathrm{T}}$. The objective (5) now becomes

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}} M \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} N \boldsymbol{\alpha}}. \tag{7}$$

Since N is positive semidefinite and singular by construction, its regularized version $N_\beta := N + \beta I$ for $\beta > 0$ is used instead. The coefficients $\boldsymbol{\alpha}$ of (7) are determined by the leading eigenvector of $(N_\beta^{-1} M)$, which is equivalent (up to scaling) to $\boldsymbol{\alpha} = N_\beta^{-1} K \mathbf{z}$. The bias becomes $b = -\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} K \mathbf{z}_+$, where $\mathbf{z}_+ := [\frac{1}{n_1} \mathbf{1}_{n_1}^{\mathrm{T}}, \frac{1}{n_2} \mathbf{1}_{n_2}^{\mathrm{T}}]^{\mathrm{T}}$. Hence, given that $\mathbf{k}_x = [k(x_1, x), \ldots, k(x_n, x)]^{\mathrm{T}}$, a two-class IKFD is defined as

$$f(x) = (\mathbf{z}^{\mathrm{T}} K N_\beta^{-1}) \mathbf{k}_x - \frac{1}{2} (\mathbf{z}^{\mathrm{T}} K N_\beta^{-1}) K \mathbf{z}_+. \tag{8}$$

Note that multiple-class problems are usually solved by one-against-all two-class discriminants.

By comparing IKFD to KFD [23], [24], we observe that the final formulations are exactly the same: The difference lies in the *definiteness* of the kernel used. This result has an important implication in practice. Independently of the definiteness of the kernel matrix, the kernel Fisher discriminant obtained by (8) is applicable to indefinite kernels and has a geometric foundation and geometric interpretation in indefinite spaces. Details on IKFD can be found in [14].

4 Kernel Quadratic Discriminant Analysis

Quadratic discriminant analysis originally assumes a finite-dimensional vectorial input space $\mathcal{X} := \mathbb{R}^k$. Each class ω_j is assumed to be normally distributed,

$$\begin{aligned} p(x|\omega_j) &= \mathcal{N}(x; \{\Sigma^{[j]}, \mu^{[j]}\}) \\ &= \frac{\exp\{-\frac{1}{2}(x - \mu^{[j]})^{\mathrm{T}} (\Sigma^{[j]})^{-1} (x - \mu^{[j]})\}}{(2\pi)^{\frac{k}{2}} (\det(\Sigma^{[j]}))^{\frac{1}{2}}}, \end{aligned}$$

with a covariance matrix $\Sigma^{[j]} \in \mathbb{R}^{k \times k}$ and a mean vector $\mu^{[j]} \in \mathbb{R}^k$. Each class has an individual prior probability $P(\omega_j)$ with $\sum_j P(\omega_j) = 1$; cf. [7] for details. The maximum a posteriori probability (MAP) decision for a pattern x relies on a comparison of c functions $p(x|\omega_j) P(\omega_j)/p(x)$, which simplify to the following quadratic discriminant functions f_j, $j = 1, \ldots, c$:

$$\begin{aligned} f_j(x) &:= -\frac{1}{2} \underbrace{(x - \mu^{[j]})^{\mathrm{T}} (\Sigma^{[j]})^{-1} (x - \mu^{[j]})}_{= D_M^2(x, \{\mu^{[j]}, \Sigma^{[j]}\})} + b_j, \\ b_j &:= -\frac{1}{2} \ln(\det(\Sigma^{[j]})) + \ln(P(\omega_j)). \end{aligned} \tag{9}$$

Given c classes, a new object x is assigned to the class ω_j if

$$f_j(x) \geq f_i(x), \quad \text{for all } i \neq j. \tag{10}$$

In case of ties, a deterministic rule is applied that, e.g., chooses minimal j that yields the maximum $f_j(x)$. In practice, covariance matrices, means, and prior probabilities are frequently unknown and estimated from the training data. In particular, the prior probabilities are usually estimated as $P(\omega_j) := n_j/n$.

As discussed in [19], nonlinear classifiers may be required in the kernel-induced feature space and Gaussian distributions can be observed. However, the authors state that, for an operator T, the term $\langle \psi(x), T\psi(x) \rangle_{\mathcal{H}}$ cannot be expressed by inner products; hence, cannot be kernelized. It is actually possible to do so if T is the empirical covariance operator, i.e., $T = C$. This is our motivation for studying quadratic classifiers based on Mahalanobis distances in the implicit kernel feature space.

Hence, in order to describe KQD, we replace x by $\phi(x)$ on the right-hand side of (9) and provide suitable approximations for the covariance operator and the mean. Most importantly, we need to find the kernel formulation of the square Mahalanobis distance. The decision rule f_j in (10) remains unchanged. The bias b_j in (9) can be expressed by operations on the kernel only, but it will get another treatment in Section 4.4. This is done in order to avoid numerical difficulties.

TABLE 2
KQD and IKQD Approaches for a c-Class Problem Based on Decision Functions f_j, $j = 1, \ldots, c$

BASIC DEFINITIONS	
$K^{[j]} = (\Psi^{[j]})^* \Psi^{[j]} \in \mathbb{R}^{n_j \times n_j}$ $K_j = \Psi^* \Psi^{[j]} \in \mathbb{R}^{n \times n_j}$	$\tilde{K}^{[j]} = H^{[j]} K^{[j]} H^{[j]} \in \mathbb{R}^{n_j \times n_j}$ $\tilde{K}_j = \tilde{\Psi}^* \tilde{\Psi}^{[j]} \in \mathbb{R}^{n \times n_j}$ $\tilde{\tilde{K}}^{[j]} = \tilde{K}_j H^{[j]} \tilde{K}_j^{\mathsf{T}} \in \mathbb{R}^{n \times n}$
$\mathbf{k}_x^{[j]} = (\Psi^{[j]})^* \psi(x) \in \mathbb{R}^{n_j \times 1}$ $\mathbf{k}_x = [k(x_1, x), \ \ldots, \ k(x_n, x)]^{\mathsf{T}}$	$\tilde{\mathbf{k}}_x^{[j]} = H^{[j]}(\mathbf{k}_x^{[j]} - \frac{1}{n_j} K^{[j]} \mathbf{1}_{n_j}) \in \mathbb{R}^{n_j \times 1}$ $\tilde{\mathbf{k}}_x = H(\mathbf{k}_x - \frac{1}{n} K \mathbf{1}_n) \in \mathbb{R}^{n \times 1}$ $\tilde{\tilde{\mathbf{k}}}_x^{[j]} = \tilde{\mathbf{k}}_x - \frac{1}{n_j} \tilde{K}_j \mathbf{1}_{n_j} \in \mathbb{R}^{n \times 1}$
$k_{xx} = \psi(x)^* \psi(x)$	$\tilde{k}_{xx}^{[j]} = k_{xx} - \frac{2}{n_j} \mathbf{1}_{n_j}^{\mathsf{T}} \mathbf{k}_x^{[j]} + \frac{1}{n_j^2} \mathbf{1}_{n_j}^{\mathsf{T}} K^{[j]} \mathbf{1}_{n_j}$

KQD AND IKQD METHODS	
KQD-IC$^+$/ IKQD-IC$^+$	**KQD-IC$^-$/ IKQD-IC$^-$**
$f_j(x) = -\frac{n_j}{2} (\tilde{\mathbf{k}}_x^{[j]})^{\mathsf{T}} (\tilde{K}_{\text{reg}}^{[j]})^{-2} \tilde{\mathbf{k}}_x^{[j]} + b_j$ $\tilde{K}_{\text{reg}}^{[j]} = \begin{cases} \tilde{K}^{[j]} + \alpha_j I_{n_j}, & \text{for KQD-IC}^+ \\ U^{[j]}(\Lambda^{[j]} + \alpha_j J^{[j]})(U^{[j]})^{\mathsf{T}}, & \text{for IKQD-IC}^+ \end{cases}$	$f_j(x) = -\frac{n_j}{2} (\tilde{\mathbf{k}}_x^{[j]})^{\mathsf{T}} ((\tilde{K}^{[j]})^-)^2 \tilde{\mathbf{k}}_x^{[j]} + b_j$ $(\tilde{K}^{[j]})^- = \text{pinv}(\tilde{K}^{[j]}, \alpha_j)$
KQD-RC$^+$/ IKQD-RC$^+$	**KQD-RC$^-$/ IKQD-RC$^-$**
$f_j(x) = -\frac{1}{2\sigma_j^2} \left(\tilde{k}_{xx}^{[j]} - (\tilde{\mathbf{k}}_x^{[j]})^{\mathsf{T}} (\tilde{K}_{\text{reg}}^{[j]})^{-1} \tilde{\mathbf{k}}_x^{[j]} \right) + b_j$ $\tilde{K}_{\text{reg}}^{[j]} = \tilde{K}^{[j]} + n_j \sigma_j^2 \begin{cases} I_{n_j}, & \text{for KQD-RC}^+ \\ U^{[j]} J^{[j]} (U^{[j]})^{\mathsf{T}}, & \text{for IKQD-RC}^+ \end{cases}$	$f_j(x) = -\frac{1}{2\sigma_j^2} \left(\tilde{k}_{xx}^{[j]} - \frac{1}{n_j \sigma_j^2} (\tilde{\mathbf{k}}_x^{[j]})^{\mathsf{T}} A \tilde{\mathbf{k}}_x^{[j]} \right) + b_j$ $A = \begin{cases} I_{n_j} & \text{for KQD-RC}^- \\ U^{[j]} J^{[j]} (U^{[j]})^{\mathsf{T}}, & \text{for IKQD-RC}^- \end{cases}$
KQD-FK$^+$/ IKQD-FK$^+$	**KQD-FK$^-$/ IKQD-FK$^-$**
$f_j(x) = -\frac{n_j}{2} (\tilde{\tilde{\mathbf{k}}}_x^{[j]})^{\mathsf{T}} (\tilde{\tilde{K}}_{\text{reg}}^{[j]})^{-1} \tilde{\tilde{\mathbf{k}}}_x^{[j]} + b_j$ $\tilde{\tilde{K}}_{\text{reg}}^{[j]} = \tilde{\tilde{K}}^{[j]} + \alpha_j \, I_n$	$f_j(x) = -\frac{n_j}{2} (\tilde{\tilde{\mathbf{k}}}_x^{[j]})^{\mathsf{T}} (\tilde{\tilde{K}}^{[j]})^- \tilde{\tilde{\mathbf{k}}}_x^{[j]} + b_j$ $(\tilde{\tilde{K}}^{[j]})^- = \text{pinv}(\tilde{\tilde{K}}^{[j]}, \alpha_j)$

The sample x is classified to ω_j iff $f_j(x) \geq f_i(x)$, for all $i \neq j$. The values b_j are found by error minimization on the training set. For simplicity, we use Ψ to denote feature spaces and * to denote the conjugate-transpose for both Hilbert and Krein spaces. Recall that $H = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}$ and $H^{[j]} = I_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^{\mathrm{T}}$. $\text{pinv}(A, \alpha)$ denotes a denoised pseudoinverse of the matrix A such that singular values whose magnitudes are smaller than α are set to zero. $\tilde{K}^{[j]} = U^{[j]} |\Lambda^{[j]}| J^{[j]} (U^{[j]})^{\mathrm{T}}$ stands for an eigendecomposition of $\tilde{K}^{[j]}$, where $\Lambda^{[j]} = \text{diag}(\boldsymbol{\lambda}_+^{[j]}, \boldsymbol{\lambda}_-^{[j]}, \mathbf{0}) = |\Lambda^{[j]}| J^{[j]}$ has p_j positive and q_j negative eigenvalues, and $J^{[j]} := \text{diag}(\mathbf{1}_{p_j}, -\mathbf{1}_{q_j}, \mathbf{1}_{n_j - p_j - q_j})$.

We will now derive three approaches to kernel quadratic discriminant denoted as: KQD-IC for *Invertible Covariance operators*, KQD-RC for *Regularized Covariance operators,* and KQD-FK for *Full Kernel matrix*. The methods differ in their underlying assumptions, the computational complexity, and the amount of the kernel-matrix information they rely on. The first two techniques work in class-related subspaces, while the third one is defined in the complete kernel-induced space. KQD-IC and KQD-RC are computationally more attractive than KQD-FK; hence, they are preferred in the case of "clean" classes, i.e., when the classes are discriminative based on the diagonal kernel submatrices. This is the reason for considering multiple formulations.

Each of the above methods has a proper extension to indefinite kernels yielding IKQD-IC, IKQD-RC, and IKQD-FK, respectively. Different regularization methods are indicated by additional subscripts and superscripts. In particular, superscript $^+$ indicates regularization by a suitable *addition*, while superscript $^-$ indicates regularization by a suitable *removal* (or *simplification*) step. All of the methods are summarized in Table 2.

4.1 KQD-IC Based on Invertible Covariance Operators

We assume an embedding of the training data by a kernel-induced mapping ϕ into a Hilbert space $\mathcal{H}$. We require here invertible (nonsingular) empirical class covariance operators C in the kernel-induced space. This limits our reasoning to a finite-dimensional $\mathcal{H}$ because the image of an empirical covariance operator C based on n samples has a finite dimension $m < n$. The following considerations require identical classwise derivations. Therefore, in order to simplify the notation, we concentrate on a single class of n elements $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ and drop the super/subscript j. Remember that the empirical mean of Φ is $\phi_\mu := \frac{1}{n} \Phi \mathbf{1}_n$, the centered configuration is $\tilde{\Phi} := \Phi - \phi_\mu \mathbf{1}_n^{\mathrm{T}} = \Phi H$, and the invertible (due to our assumption) empirical covariance operator is defined as $C := \frac{1}{n} \tilde{\Phi} \tilde{\Phi}^{\mathrm{T}}$. We want to kernelize the empirical square Mahalanobis distance $D_M^2(\phi(x); \{\phi_\mu, C\})$ given in (1). This can be computed without performing the explicit mapping ϕ as we will now derive. Similar derivations for the subsequent methods are presented in Appendix A, which can be found in the Computer Society Digital library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290.

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 given by $\tilde{\Phi} = USV^{\mathrm{T}}$ with orthogonal matrices $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$, and a diagonal matrix $S \in \mathbb{R}^{m \times n}$. By using the orthogonality of U and V, we have: $C = \frac{1}{n} \tilde{\Phi} \tilde{\Phi}^{\mathrm{T}} = \frac{1}{n} U S S^{\mathrm{T}} U^{\mathrm{T}}$ and $K = \tilde{\Phi}^{\mathrm{T}} \tilde{\Phi} = V S^{\mathrm{T}} S V^{\mathrm{T}}$, with an invertible matrix $SS^{\mathrm{T}} \in \mathbb{R}^{m \times m}$, but singular $S^{\mathrm{T}} S \in \mathbb{R}^{n \times n}$. So, $C^{-1} = nU(SS^{\mathrm{T}})^{-1} U^{\mathrm{T}}$ and $\tilde{K}^- = V(S^{\mathrm{T}} S)^- V^{\mathrm{T}}$, where the superscript $^-$ denotes here the pseudoinverse of a matrix. This

Moore-Penrose inverse exists, is unique, and can be obtained by a singular value decomposition of the matrix. (Specific conditions for the existence of the Moore-Penrose inverse can also be characterized for general operators in Krein spaces; see [22] for details.) Multiplication of these equations with $\tilde{\Phi}$ yields

$$\begin{aligned} \frac{1}{n} C^{-1}\tilde{\Phi} &= U(SS^{\mathrm{T}})^{-1}SV^{\mathrm{T}}, \\ \tilde{\Phi}\tilde{K}^{-} &= US(S^{\mathrm{T}}S)^{-}V^{\mathrm{T}}. \end{aligned} \tag{11}$$

Since $S \in \mathbb{R}^{m\times n}$ is diagonal and has m nonzero singular values, both middle matrices $S(S^{\mathrm{T}}S)^{-}$ and $(SS^{\mathrm{T}})^{-1}S$ are $m \times n$ diagonal matrices with inverted singular values on the diagonal. Therefore, these matrices are identical and we conclude that

$$\tilde{\Phi}\tilde{K}^{-} = \frac{1}{n} C^{-1}\tilde{\Phi}. \tag{12}$$

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

$$\begin{aligned} C\tilde{\phi}(x) &= \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^{\mathrm{T}}\left(\phi(x) - \frac{1}{n}\Phi\mathbf{1}_n\right) \\ &= \frac{1}{n}\tilde{\Phi}H\left(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n\right) = \frac{1}{n}\tilde{\Phi}\tilde{\mathbf{k}}_x, \end{aligned}$$

where $\mathbf{k}_x$ and $\tilde{\mathbf{k}}_x$ are defined in (2). Since C is invertible, this implies with (12) that

$$\tilde{\phi}(x) = \frac{1}{n} C^{-1}\tilde{\Phi}\tilde{\mathbf{k}}_x = \tilde{\Phi}\tilde{K}^{-}\tilde{\mathbf{k}}_x. \tag{13}$$

Finally, the identities (12) and (13) allow us to express the Mahalanobis distance in its kernelized form as

$$\begin{aligned} D_M^2(\phi(x); \{\phi_\mu, C\}) &= \tilde{\phi}(x)^{\mathrm{T}} C^{-1}\tilde{\phi}(x) = \tilde{\phi}(x)^{\mathrm{T}} C^{-1}\tilde{\Phi}\tilde{K}^{-}\tilde{\mathbf{k}}_x \\ &= n\,\tilde{\mathbf{k}}_x^{\mathrm{T}}(\tilde{K}^{-})^2\tilde{\mathbf{k}}_x. \end{aligned}$$

For the jth class, the kernel Mahalanobis distance becomes:

$$D_M^2(\phi(x); \{\phi_\mu^{[j]}, C^{[j]}\}) = n_j(\tilde{\mathbf{k}}_x^{[j]})^{\mathrm{T}}((\tilde{K}^{[j]})^{-})^2\tilde{\mathbf{k}}_x^{[j]},$$

where $\tilde{K}^{[j]} = H^{[j]}K^{[j]}H^{[j]}$ and $\tilde{\mathbf{k}}_x^{[j]} = H^{[j]}(\mathbf{k}_x^{[j]} - \frac{1}{n_j}K^{[j]}\mathbf{1}_{n_j})$. In a c-class problem, a quadratic discriminant for the jth class is obtained from (9) by inserting the estimated quantities as

$$f_j(x) = -\frac{n_j}{2}(\tilde{\mathbf{k}}_x^{[j]})^{\mathrm{T}}((\tilde{K}^{[j]})^{-})^2\tilde{\mathbf{k}}_x^{[j]} + b_j. \tag{14}$$

Note that $\tilde{K}^{[j]}$ is singular as $\mathrm{rank}(\tilde{K}^{[j]}) < n_j$ due to kernel centering. In addition, we can rely on the pseudoinverse of $\tilde{K}^{[j]}$ with a given tolerance $\alpha_j > 0$, such that $(\tilde{K}^{[j]})^{-} := \mathrm{pinv}(\tilde{K}^{[j]}, \alpha_j)$. This means that singular values smaller than α_j are treated as 0. The use of the tolerance α_j acts as a denoising step. It is necessary in practical applications in order to prevent noisy and unreliable estimates of $(\tilde{K}^{[j]})^{-}$ when $\tilde{K}^{[j]}$ yields many tiny eigenvalues. Alternatively, we can use the inverse of the regularized kernel $\tilde{K}_{\mathrm{reg}}^{[j]} := \tilde{K}^{[j]} + \alpha_j I_{n_j}$, where $\alpha_j > 0$ is a small regularization constant. This leads to alternative discriminant functions in the form of

$$f_j(x) = -\frac{n_j}{2}(\tilde{\mathbf{k}}_x^{[j]})^{\mathrm{T}}(\tilde{K}_{\mathrm{reg}}^{[j]})^{-2}\tilde{\mathbf{k}}_x^{[j]} + b_j. \tag{15}$$

We will denote method (15) by $\mathrm{KQD-IC}^{+}$(KQD with Invertible Covariance matrices), while method (14) is denoted by KQD-IC^{-}. Here, the superscript $^{+}$ indicates regularization by *diagonal addition*, while $^{-}$ indicates *removal* of kernel matrix information by a thresholded pseudoinverse.

4.1.1 IKQD-IC, Extension to Indefinite Kernels

We again assume nonsingular empirical class covariance operators of data embedded into a finite-dimensional Krein space $\mathcal{K}$. One can show with a slightly refined argumentation that the analogue of (12) also holds for the indefinite case. Hence, we can express the kernel Mahalanobis distance as before. We omit the derivation here and refer to Appendix A.1, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290, for details. As a result, we obtain the following quadratic discriminants for the IKQD-IC^{-}approach:

$$f_j(x) = -\frac{n_j}{2}(\tilde{\mathbf{k}}_x^{[j]})^{\mathrm{T}}((\tilde{K}^{[j]})^{-})^2\tilde{\mathbf{k}}_x^{[j]} + b_j, \tag{16}$$

where $(\tilde{K}^{[j]})^{-} := \mathrm{pinv}(\tilde{K}^{[j]}, \alpha_j)$ is a denoised pseudoinverse of $\tilde{K}^{[j]}$. It means that singular values whose *magnitudes* are smaller than the chosen α_j are set to zero. This formulation is equivalent to (14) except for the definiteness of the kernel matrix $\tilde{K}^{[j]}$. The inverse of regularized $\tilde{K}^{[j]}$ can again be used instead of the pseudoinverse $(\tilde{K}^{[j]})^{-}$. Remember that, when we regularize $\tilde{K}^{[j]}$ by adding a constant α_j to its diagonal, $\tilde{K}^{[j]} + \alpha_j I_{n_j}$, we equivalently enlarge the original eigenvalues of $\tilde{K}^{[j]}$ by α_j. Here, $\tilde{K}^{[j]}$ is an indefinite kernel that has both positive and negative eigenvalues. Regularization should therefore be in agreement with this property. Let us consider an eigendecomposition of $\tilde{K}^{[j]}$ as $\tilde{K}^{[j]} = U^{[j]}\Lambda^{[j]}(U^{[j]})^{\mathrm{T}}$, where $\Lambda^{[j]} = \mathrm{diag}(\boldsymbol{\lambda}_+^{[j]}, \boldsymbol{\lambda}_-^{[j]}, \mathbf{0})$ is a diagonal matrix with p_j positive, q_j negative, and $(n_j - p_j - q_j)$ zero eigenvalues, while the corresponding eigenvectors are stored in $U^{[j]} = [U_+^{[j]}, U_-^{[j]}, U_0^{[j]}]$. By introducing $J^{[j]} := \mathrm{diag}(\mathbf{1}_{p_j}, -\mathbf{1}_{q_j}, \mathbf{1}_{n_j-p_j-q_j})$, we imply that $\Lambda^{[j]} = |\Lambda^{[j]}|J^{[j]}$, where $|\Lambda^{[j]}|$ denotes the absolute values of $\Lambda^{[j]}$. We can then easily verify that $\tilde{K}^{[j]} = U^{[j]}|\Lambda^{[j]}|J^{[j]}(U^{[j]})^{\mathrm{T}}$. Hence, we will define $\tilde{K}_{\mathrm{reg}}^{[j]} := U^{[j]}\Lambda_{\mathrm{reg}}^{[j]}(U^{[j]})^{\mathrm{T}}$, where $\Lambda_{\mathrm{reg}}^{[j]} := \Lambda^{[j]} + \alpha_j J^{[j]}$ and $\alpha_j > 0$ is a chosen constant. This leads to the IKQD-IC^{+}discriminants in the following form:

$$f_j(x) = -\frac{n_j}{2}(\tilde{\mathbf{k}}_x^{[j]})^{\mathrm{T}}(\tilde{K}_{\mathrm{reg}}^{[j]})^{-2}\tilde{\mathbf{k}}_x^{[j]} + b_j. \tag{17}$$

These are equivalent to (15) when $\tilde{K}^{[j]}$ is a pd kernel matrix.

4.2 KQD-RC Based on Regularized Covariance Operators

Since we deal with finite samples in a high-dimensional or infinite dimensional Hilbert space $\mathcal{H}$, the empirical covariance operator may not be invertible. Regularization is therefore necessary to prevent it from being singular. One can show that an additive regularization of the covariance operator $C_{\mathrm{reg}}^{[j]} := \frac{1}{n_j}\tilde{\Phi}^{[j]}(\tilde{\Phi}^{[j]})^{\mathrm{T}} + \sigma_j^2 I$ is equivalent to an additive regularization of the centered kernel matrix

$\tilde{K}^{[j]}_{\text{reg}} := \tilde{K}^{[j]} + n_j\sigma_j^2 I_{n_j}$. This allows subsequent derivation of the corresponding kernel Mahalanobis distance. See Appendix A.2, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290, for details.

In our c-class problem, a quadratic discriminant for the jth class described by $\{\phi^{[j]}_\mu, C^{[j]}_{\text{reg}}\}$ is therefore defined as

$$f_j(x) = -\frac{1}{2\sigma_j^2}\left(\tilde{k}^{[j]}_{xx} - (\tilde{\mathbf{k}}^{[j]}_x)^{\mathrm{T}}(\tilde{K}^{[j]}_{\text{reg}})^{-1}\tilde{\mathbf{k}}^{[j]}_x\right) + b_j. \quad (18)$$

We refer to this method as KQD-RC$^+$ (Kernel Quadratic Discriminant with Regularized Covariance operators). There is no need to use a pseudoinverse here as $\tilde{K}^{[j]}_{\text{reg}}$ is invertible. Note, instead, that $n_j\sigma_j^2 I_{n_j}$ is a dominant component in $\tilde{K}^{[j]}_{\text{reg}}$ for a sufficiently large n_j. In such a case, $(\tilde{K}^{[j]}_{\text{reg}})^{-1}$ can be approximated by $\frac{1}{n_j\sigma_j^2} I_{n_j}$. This leads to the following simplified discriminants, denoted by KQD-RC$^-$:

$$f_j(x) = -\frac{1}{2\sigma_j^2}\left(\tilde{k}^{[j]}_{xx} - \frac{(\tilde{\mathbf{k}}^{[j]}_x)^{\mathrm{T}}\tilde{\mathbf{k}}^{[j]}_x}{n_j\sigma_j^2}\right) + b_j. \quad (19)$$

4.2.1 IKQD-RC, Extension to Indefinite Kernels

Similarly to the positive definite case, we deal with finite samples in a high-dimensional or infinite-dimensional Krein space $\mathcal{K}$. So, regularization of the empirical covariance operator is necessary to prevent it from being singular. Here, however, the regularization should respect the indefinite character of the space, i.e., be in agreement with the positive and negative subspaces of $\mathcal{K}$. The derivations in Appendix A.3, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290, are based on the choice $\tilde{K}^{[j]}_{\text{reg}} := \tilde{K}^{[j]} + n_j\sigma_j^2 U^{[j]}J^{[j]}(U^{[j]})^{\mathrm{T}}$, where $\tilde{K}^{[j]} = U^{[j]}\Lambda^{[j]}(U^{[j]})^{\mathrm{T}}$ is the eigendecomposition of the centered kernel submatrix for the jth class and $J^{[j]} := \text{diag}(\mathbf{1}_{p_j}, -\mathbf{1}_{q_j}, \mathbf{1}_{n_j-p_j-q_j})$ with p_j and q_j being the number of positive and negative eigenvalues of $\Lambda^{[j]}$, respectively. This leads to the kernel Mahalanobis distance and allows us to define a quadratic discriminant for the jth class as

$$f_j(x) = -\frac{1}{2\sigma_j^2}\left(\tilde{k}^{[j]}_{xx} - (\tilde{\mathbf{k}}^{[j]}_x)^{\mathrm{T}}(\tilde{K}^{[j]}_{\text{reg}})^{-1}\tilde{\mathbf{k}}^{[j]}_x\right) + b_j. \quad (20)$$

Note that the above expression is the same as (18), except that $\tilde{K}^{[j]}$ is now an indefinite kernel matrix and $\tilde{K}^{[j]}_{\text{reg}}$ is regularized in agreement with the indefinite character of the kernel. We will denote this method as IKQD-RC$^+$. If $n_j\sigma_j^2$ is dominating the terms in $\tilde{K}$, we can simplify this method further on by approximating $\tilde{K}^{[j]}_{\text{reg}}$ by $n_j\sigma_j^2 U^{[j]}J^{[j]}(U^{[j]})^{\mathrm{T}}$. Hence, $(\tilde{K}^{[j]}_{\text{reg}})^{-1} = \frac{1}{n_j\sigma_j^2}U^{[j]}J^{[j]}(U^{[j]})^{\mathrm{T}}$. This leads to the following IKQD-RC$^-$ discriminants:

$$f_j(x) = -\frac{1}{2\sigma_j^2}\left(\tilde{k}^{[j]}_{xx} - \frac{(\tilde{\mathbf{k}}^{[j]}_x)^{\mathrm{T}}U^{[j]}J^{[j]}(U^{[j]})^{\mathrm{T}}\tilde{\mathbf{k}}^{[j]}_x}{n_j\sigma_j^2}\right) + b_j. \quad (21)$$

Note that $U^{[j]}J^{[j]}(U^{[j]})^{\mathrm{T}}$ is *not* a diagonal matrix, in contrast to KQD-RC$^-$ for which $J^{[j]} = I_{n_j}$ holds.

4.3 KQD-FK Derived in the Complete Kernel Space

Both KQD approaches considered so far build discriminant functions f_j in class-related kernel feature subspaces. The functions f_j rely on the $n_j \times n_j$ class kernel matrices $K^{[j]}$, which are diagonal block submatrices of the kernel K. This means that the between-class information expressed in the relevant off-diagonal $(n-n_j)\times n_j$ submatrices of the kernel K is unused in the Mahalanobis distances. Now, we want to propose the third approach, in which each discriminant function f_j relies on both within-class and between-class kernel information. As a result, it builds upon kernel values from the objects of the jth class to all other objects. We therefore define KQD in a complete kernel space specified via kernel PCA (KPCA), as derived in Appendix A.4, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290.

Let $\tilde{K} = [\tilde{K}_1, \ldots, \tilde{K}_c]$ be the centered kernel matrix, where centering is global for all training objects. The column-blocks $\tilde{K}_j \in \mathbb{R}^{n\times n_j}$ correspond to the kernel vectors of different classes. The lower subscript is chosen to avoid confusion with the classwise centered matrices $\tilde{K}^{[j]} \in \mathbb{R}^{n_j\times n_j}$ from the KQD-IC methods. The kernelized Mahalanobis distance is based on the matrix $\tilde{\tilde{K}}^{[j]} := \tilde{K}_j H^{[j]}\tilde{K}_j^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. Note that $\tilde{K}_j$ is a submatrix of $\tilde{K}$, where $\tilde{K}$ is centered as a whole, while $\tilde{\tilde{K}}^{[j]}$ is an inner product matrix that involves additional centering of $\tilde{K}_j$ with respect to the jth class. For this reason, we use the double-tilde notation. Since $\text{rank}(\tilde{\tilde{K}}^{[j]}) < n_j$ by construction, its inverse cannot be derived. In analogy to KQD-IC, we either use a pseudoinverse of $\tilde{\tilde{K}}^{[j]}$ or regularize it by diagonal addition. This leads to the following discriminant functions, denoted as KQD-FK$^-$(Kernel Quadratic Discriminant in the Full Kernel space)

$$f_j(x) := -\frac{n_j}{2}(\tilde{\tilde{\mathbf{k}}}^{[j]}_x)^{\mathrm{T}}(\tilde{\tilde{K}}^{[j]})^{-}\tilde{\tilde{\mathbf{k}}}^{[j]}_x + b_j, \quad (22)$$

with $\tilde{\tilde{\mathbf{k}}}^{[j]}_x := \tilde{\mathbf{k}}_x - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}$. The KQD-FK$^+$approach is based on $\tilde{\tilde{K}}^{[j]}_{\text{reg}} := \tilde{\tilde{K}}^{[j]} + \alpha_j I_n$, which leads to the following discriminants:

$$f_j(x) := -\frac{n_j}{2}(\tilde{\tilde{\mathbf{k}}}^{[j]}_x)^{\mathrm{T}}(\tilde{\tilde{K}}^{[j]}_{\text{reg}})^{-1}\tilde{\tilde{\mathbf{k}}}^{[j]}_x + b_j. \quad (23)$$

4.3.1 IKQD-FK, Extension to Indefinite Kernels

We denote $\tilde{K} = [\tilde{K}_1, \ldots, \tilde{K}_c]$ as the centered indefinite kernel matrix for all training objects. The column-blocks $\tilde{K}_j \in \mathbb{R}^{n\times n_j}$ describe kernel vectors of different classes. A data representation obtained from indefinite kernel PCA (IKPCA) [27] allows one to derive the kernel Mahalanobis distance based on the double-centered matrix $\tilde{\tilde{K}}^{[j]} := \tilde{K}_j H^{[j]}\tilde{K}_j^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ as worked out in Appendix A.5, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.290. Again, $\text{rank}(\tilde{\tilde{K}}^{[j]}) < n_j$ by construction, so its inverse

cannot be computed. We can either use a pseudoinverse of $\tilde{\tilde{K}}^{[j]}$ or regularize it appropriately. Note that independent of the definiteness of $\tilde{K}$, $\tilde{\tilde{K}}^{[j]}$ is always positive semidefinite because it is an inner product matrix, i.e., $\tilde{\tilde{K}}^{[j]} = (\tilde{K}_j H^{[j]})(\tilde{K}_j H^{[j]})^{\mathrm{T}}$. Consequently, both pseudoinverse of $\tilde{\tilde{K}}^{[j]}$ and its regularization by diagonal addition work identically to the positive definite case. This means that the IKQD-FK$^-$ discriminants are described by (22), while the IKQD-FK$^+$ discriminants are expressed by (23). The difference again only lies in the definiteness of the kernel. The summary of all KQD approaches is presented in Table 2.

4.4 Choice of Bias

It is possible to derive the bias b_j in the discriminant function f_j of an MAP decision only by using operations on kernels. For instance, we get $b_j = -\frac{1}{2}\sum_{i=1}^{l} \ln(\lambda_i^{[j]}) + \ln(P(\omega_j))$, where $\lambda_i^{[j]}$ are nonzero eigenvalues of a nonsingular covariance matrix $C^{[j]}$ in an l-dimensional space. This holds because $\ln(\det(C^{[j]})) = \ln(\prod_{i=1}^{l} \lambda_i^{[j]}) = \sum_{i=1}^{l} \ln(\lambda_i^{[j]})$. It is well known, e.g., from KPCA, that $\lambda_i^{[j]}$ can be obtained as the l nonzero eigenvalues of the scaled and centered kernel matrix $\frac{1}{n}\tilde{K}^{[j]}$, cf. [33]. In particular, it is straightforward to show that the eigenvalues $\lambda_i^{[j]}$ of $C^{[j]}$ are identical to the eigenvalues of $\frac{1}{n_j}\tilde{K}^{[j]}$ for $i = 1, \ldots, l := \mathrm{rank}(\tilde{K}^{[j]})$. Similar expressions for b_j can also be derived for regularized covariance matrices.

Numerical problems, however, arise because a centered kernel matrix has often a slowly decaying eigenvalue spectrum. In order to take all nonzero eigenvalues into account, one has to compute the logarithm of the eigenvalue-product. This is numerically unstable if $\tilde{K}^{[j]}$ has many small eigenvalues. The restriction to a fixed number of eigenvalues is equivalent to the choice of intrinsic dimension. A variation of this factor can lead to large variations in the bias as the logarithms of small eigenvalues become arbitrarily large in magnitude. In addition to the instability of a proper estimation of intrinsic dimension, the resulting (unreliable) bias b_j turns out to frequently dominate the Mahalanobis distance contribution in the experimental computation of f_j. As a result, it spoils the predictability of the resulting classification rule. Therefore, we apply another interpretation of the bias values b_j as in the traditional QDA. This leads to a stable and elegant computation scheme for b_j in the kernelized classifiers.

In the case of classwise normally distributed data, the traditional QDA (with exact mean and covariance) is the Bayes classifier [7]. In particular, no other choice of bias values will result in a lower classification error than the Bayes error. Therefore, the bias values of f_j can equivalently be defined as the ones that minimize the QDA prediction error. Since the training error is a good surrogate for the Bayes error in QDA for a large training set, we apply the following procedure to determine b_j on the training data. For a two-class problem, say ω_i and ω_j, a greedy search can be applied to determine the optimal estimate for the biases b_i and b_j, or more precisely, for their difference $b_i - b_j$. This difference is the only relevant quantity for the class decisions, as an addition of a constant to all bias values keeps the decisions unchanged. Having fixed one value b_i, only a finite number of values for the second bias b_j need to be tried to obtain the minimal training error for the two classes. For a c-class problem, this can be applied in a classwise manner which yields $\frac{1}{2}c(c-1)$ estimates Δ_{ij} for the differences $b_i - b_j$, $j > i$. The desired bias values $\mathbf{b} = [b_i]_{i=1}^{c}$ are found by solving a small least squares problem

$$\min_{\mathbf{b}} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} (b_i - b_j - \Delta_{ij})^2.$$

TABLE 3
Train and Test Complexity for Different Classifiers

Method	Train complexity	Test complexity
KQD-IC/ IKQD-IC	$\mathcal{O}(n^3/c + n^2 + c^3)$	$\mathcal{O}(n^2/c)$
KQD-RC$^+$/ IKQD-RC$^+$	$\mathcal{O}(n^3/c + n^2 + c^3)$	$\mathcal{O}(n^2/c)$
KQD-RC$^-$	$\mathcal{O}(n^2 + c^3)$	$\mathcal{O}(n/c)$
IKQD-RC$^-$	$\mathcal{O}(n^3/c + n^2 + c^3)$	$\mathcal{O}(n^2/c)$
KQD-FK/ IKQD-FK	$\mathcal{O}(cn^3 + c^3)$	$\mathcal{O}(cn^2)$
KFD/ IKFD (one-vs-all)	$\mathcal{O}(cn^3)$	$\mathcal{O}(cn)$
SVM /ISVM (one-vs-all)	$\mathcal{O}(cn^2)$	$\mathcal{O}(cn\nu)$
KNN / IKNN	$-$	$\mathcal{O}(kn)$
KPCA-QD / IKPCA-QD	$\mathcal{O}(pn^3 + cp^3)$	$\mathcal{O}(np + cp^2)$

4.5 Computational Complexity

Table 3 presents computational complexities of the KQD approaches and some reference methods. The latter are linear kernel classifiers, such as KFD and SVM, and nonlinear ones, such as the kernel k-Nearest-Neighbor (KNN) based on the kernel-induced distance $d^2(x, x') := k(x, x) - 2k(x, x') + k(x', x')$, and KPCA-QD, which is a quadratic discriminant trained in a feature space obtained from KPCA. Indefinite versions of these are identified by the prefix "I" in the used abbreviations.

For simplicity, we assume a c-class problem with n training samples such that class priors are equal and set to $n_j := n/c$ for all classes. The value ν denotes the fraction of support vectors of SVM, while p is the dimension of the KPCA space. The test complexities of the KQD methods rely on c evaluations of decision functions. These are either matrix-vector multiplications of size n_j for classwise methods (except for KQD-RC$^-$) of size n for full kernel approaches, or merely vector inner products of the length n_j for KQD-RC$^-$. This leads to the test complexities for a single pattern reported in the right column of Table 3. The bias derivation for the classifiers requires cn Mahalanobis distances, equal to n times the test-complexity. For these values, the bias-difference estimates Δ_{ij} are computed in $\mathcal{O}(c^2 n_j^2) = \mathcal{O}(n^2)$ and the solution of a least square problem finds the desired bias values in $\mathcal{O}(c^3)$, as described in Section 4.4. The computation of c matrix (pseudo)inverses for the decision functions can be realized in either $\mathcal{O}(cn_j^3)$ or $\mathcal{O}(cn^3)$, depending on the size of the involved matrices. Again, KQD-RC$^-$ is a special case as only auxiliary vectors need to be computed here in $\mathcal{O}(cn_j^2)$. This gives the training complexities as shown in the left column of Table 3. Note that centering of a kernel vector can be realized in $\mathcal{O}(n_j)$ or

$\mathcal{O}(n)$ depending on the vector length, so it does not influence the estimated values. Reference classifiers are based on matrix inverses of the complexity $\mathcal{O}(n^3)$ for KFD, eigendecomposition complexity $\mathcal{O}(n^3)$ for KPCA-QD, and empirical SVM complexity scaling with $\mathcal{O}(n^2)$ (based on optimized training routines; otherwise, the complexity of $\mathcal{O}(n^\alpha)$, $\alpha \geq 3$, would be realistic for general QP solvers).

Observe that the KQD-IC and KQD-RC approaches are clearly beneficial in the case of multiple classes as the dominating n^3-term is mainly inversely proportional to c. The more expensive KQD-FK approaches still have identical training complexity as, e.g., KFD. As we have quadratic classifiers, the test complexity is based on nonsparse matrix multiplications; hence, asymptotically more expensive than in case of linear kernel classifiers such as SVM and KFD. The KQD-RC$^-$ approach is clearly advantageous over the remaining classifiers due to its simple classification rule.

4.6 Related Methods

Various nonlinear kernel techniques, including the kernel Mahalanobis distance, are considered in [32]. As such, the pure kernel Mahalanobis distance (KMD) is used there in a classwise manner. This is analogous to our KQD-IC$^-$ approach relying on the pseudoinverse of the class-related kernel submatrices, but without the use of bias values. The authors report a good performance of KMD on the RBF kernels defined for some standard vectorial data from the Machine Learning Repository [17].

The assumption of Gaussian distributions in kernel spaces suggests a relation to Gaussian processes, i.e., collections of random variables whose any finite number has a joint Gaussian distribution. Indeed, there is an interesting link between KQD-RC$^+$ and Gaussian process regression [29]. A Gaussian process is used in Machine Learning as a prior probability distribution over functions and used for Bayesian inference. In our case, these functions are defined in a centered kernel-induced space as $f(x) = w^{\mathrm{T}}\phi(x)$. In practice, we also assume additive iid Gaussian noise ϵ with variance α_n^2, which leads to the relation $y = f(x) + \epsilon$. As a result, $f(x)$ is a Gaussian process with mean $m(x) = 0$ and covariance $k(x, x')$. Given the training data $\{x_i, y_i\}_{i=1}^n$, a centered kernel matrix $\tilde{K}$ is the covariance matrix of the corresponding Gaussian process. The joint distribution of the observed target values and the function value f_x at a test point x is

$$\begin{bmatrix} \mathbf{y} \\ f_x \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tilde{K} + \alpha_n^2 I_n & \tilde{\mathbf{k}}_x \\ \tilde{\mathbf{k}}_x & \tilde{k}_{xx} \end{bmatrix}\right).$$

The Gaussian posterior distribution $p(f_x|X_{\mathrm{tr}}, \mathbf{y}, x)$ has the mean $\bar{f}_x = \tilde{\mathbf{k}}_x^{\mathrm{T}}(\tilde{K} + \alpha_n^2 I_n)^{-1}\mathbf{y}$ and the variance $\mathrm{var}(f_x) = \tilde{k}_{xx} - \tilde{\mathbf{k}}_x^{\mathrm{T}}(\tilde{K} + \alpha_n^2 I_n)^{-1}\tilde{\mathbf{k}}_x$ (see [29] for details). Hence, in particular, $\mathrm{var}(f_x)$ with $\alpha_n^2 = n\sigma_n^2$ is equivalent to the kernel Mahalanobis distance derived for KQD-RC$^+$, i.e., when the covariance operator is regularized in the kernel-induced space.

Kernel discriminant analysis is more specifically discussed in [19]. In particular, the authors present a statistical support for KFD and show an approach of "kernel Fisher's quadratic discriminant analysis." This method uses a vectorial representation of patterns by the kernel values $\mathbf{k}_x$ and performs QDA on them. Our approaches are quite different and do not restrict the kernels to be Gaussian or Epanechnikov as considered in [19].

We also want to mention the recent paper of Wang et al. [41] which presents a method of kernel quadratic discriminant analysis for small sample size problems. This proposal relies on supervised dimension reduction in a kernel-induced space followed by discriminant analysis. Given c classes, the first step is realized by a kernel Fisher mapping to at most a $(c-1)$-dimensional space in which a specifically regularized quadratic discriminant is found. As such, this procedure is not a pure extension of the traditional quadratic discriminant, as we develop it in this paper, but involves an intermediate step of a kernel linear discriminant.

Finally, we want to emphasize that kernel methods are mostly nonlinear extensions of linear algorithms and one might ask whether KQD techniques can be interpreted as linear classification in an extended kernel-induced space with a suitably chosen kernel. The answer is negative, which can be most obviously seen in the KQD-RC approaches, as the diagonal kernel values $k(x, x)$ are required. No linear classifier in kernel space could make use of these for classification.

5 Experiments and Results

In our experimental study, we focus on various classification problems in order to compare the performance of the KQD and IKQD methods to relevant reference classifiers, such as SVM, KFD, KNN, and KPCA-QD as introduced in Section 4.5. The reference methods are also applicable to indefinite kernels, cf. [11], [27], and Section 3.1. Consequently, the reference methods will be denoted as ISVM, IKFD, IKNN and IKPCA-QD in the case of indefinite kernel matrices. All experiments rely on the MATLAB package PRtools41 [8]. SVM/ISVM is trained by using MATLAB inherent optimization routines for small data sets and LIBSVM [3] for large data sets. In particular, the latter software is guaranteed to converge for indefinite kernels.

5.1 Positive Definite Kernel on 2D Data

Let us consider an artificial data set as illustrated in Fig. 1. The classes are generated by two normal distributions, slightly transformed in a nonlinear way such that the resulting distributions are no longer Gaussian. Each class in the training set is represented by 50 samples. We choose the Gaussian Radial Basis Function (RBF), $k(x, x') = \exp(-|x - x'|^2/s^2)$, as the kernel. The same regularization parameters are used for all classes, i.e., $\sigma_j^2 := \sigma^2, \alpha_j := \alpha$, and we perform 10-fold cross-validation to determine the following parameters: $\alpha \in [10^{-10}, 10^{-3}]$ for the KQD-IC and KQD-FK methods, $\sigma^2 \in [10^{-3}, 10^4]$ for the KQD-RC approaches, $\alpha \in [10^{-6}, 10^1]$ for KFD, $C \in [10^{-1}, 10^6]$ for SVM, and $\alpha \in [10^{-7}, 10^0]$ for KPCA-QD, where each parameter interval is discretized by 15 values on a logarithmic scale. The value $k \in \{1, \ldots, 15\}$ is optimized for KNN. The kernel parameter is included in the cross-validation search by 15 values for s spanning the interval $[0.1, 500]$. Classification results are found on independently drawn test sets of $500 + 500$ examples.

Example KQD-classifiers are depicted in Fig. 1a. Fig. 1b illustrates the reference classifiers: KFD, SVM, KPCA-QD,

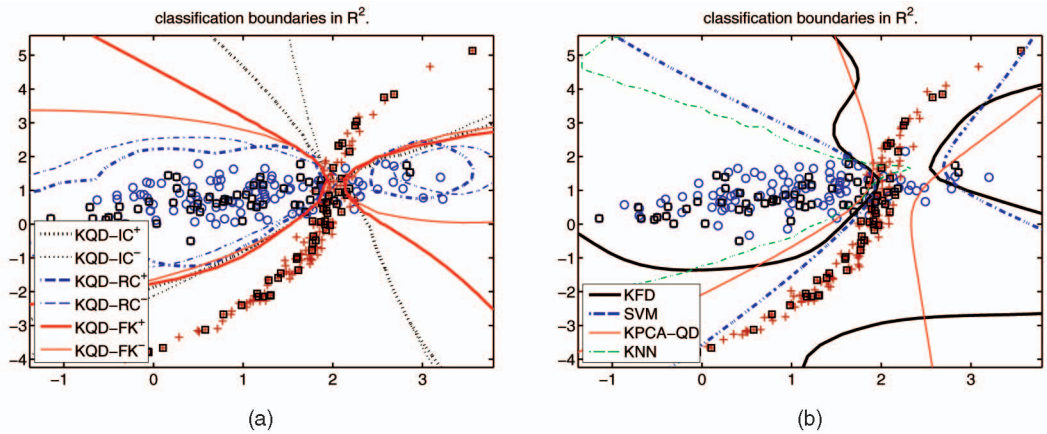


Fig. 1. Classifiers on 2D toy data based on the Gaussian RBF-kernel. The classifier parameters are determined via cross-validation. (a) Results for all KQD methods. (b) Results for the reference methods.

and KNN. The training samples are marked by squares, and additionally, a random subset of 200 test samples is plotted. Note that the cross-validated s are reflected in the variability of the decision lines, e.g., higher variability or lower s for KQD-RC$^+$ and KQD-RC$^-$. The KNN rule is, as expected, highly nonlinear.

To analyze classification performance, we determine the overall mean and standard deviation of the test errors determined by $(s, C, \alpha, \sigma, k)$ cross-validated classifiers in 10 runs. The corresponding errors are reported in the left part of Table 4. The KQD-IC and KQD-FK nonlinear kernel classifiers seem to perform much better than the SVM and KFD, which are linear classifiers in $\mathcal{H}$. But they are also superior to other nonlinear kernel classifiers, namely KNN and KPCA-QD.

5.2 Indefinite Kernel on 2D Data

We consider an artificial 4×4 checkerboard data based on a uniform distribution on $[-2, 2]^2 \subset \mathbb{R}^2$, cf. Fig. 2. We first define the base kernel $k(x, x') := \exp(-d(x, x')^4/s^2)$ by using $d(x, x') := \sum_{i=1,2} |x_i - x'_i|^2$. Practical source of indefiniteness in kernels can be caused by incorporation of prior knowledge about invariance into kernels, deriving kernels from distances, or combining kernels [26], [12]. We observe that the checkerboard distribution is invariant with respect to the point reflection $\tau(x) := -x$ through the origin. We incorporate this knowledge by combining two base kernels into a new one: $\bar{k}(x, x') := \max\{k(x, x'), k(x, \tau(x'))\}$, which can alternatively be motivated by invariant distances [13]. We choose these kernel settings because of significant indefiniteness. Hence, the example is suitable for demonstrating the behavior of the methods for indefinite kernels. Note that this kernel is symmetric, as $k(x, \tau(x')) = k(\tau(x), x')$ for the RBF-kernel.

We follow the same experimental setup as in Section 5.1, i.e., a training set of $50 + 50$ elements is drawn, kernel width and classifier parameters are found via 10-fold cross-validation. Test error rates are determined on an independent test set of $500 + 500$ samples. The ranges of s and α are slightly shifted as compared to the previous section.

Example classifiers are illustrated in Fig. 2. One can clearly observe the perfect point symmetry of all classifiers due to the use of an invariant kernel, even though the training set is asymmetric. To maintain the clarity of presentation, the test examples are not plotted.

To assess the statistical significance, we repeat the above data-drawing, cross-validation, and test-error determination 10 times. The resulting average test errors are given in the right column of Table 4 in the previous section. We see that all IKQD approaches outperform both ISVM and IKNN; IKQD-IC$^-$ is even slightly superior to IKPCA-QD and IKFD.

In the above experiments, the kernel parameter s was cross-validated, which is necessary for evaluating the classification performance. Still, further interesting observations can be made by fixing s and performing cross-validation over the remaining parameters. By this, the inherent feature-space representation of the data is fixed, which allows investigations of indefiniteness. Further preferences of s of the different classifiers can be found.

These results are presented in Table 5. In addition to the 10-fold averaged test errors, we assess some measures of indefiniteness of the resulting kernel matrices. First, we determine the signature (p, q) of the kernel matrix, defining the dimensions $p, q \in \mathbb{N}$ of positive and negative subspaces, respectively. It results from an embedding of the training data into a finite-dimensional Krein space $\mathcal{K}$. Further, we provide an index of indefiniteness, $r_{\text{neg}} := (\sum_{\lambda_i<0} |\lambda_i|)/(\sum_i |\lambda_i|)$, the ratio of negative variance to overall variance

TABLE 4
Average Classification Errors (%) for Positive Definite and Indefinite 2D Data Sets

Classifier	Positive definite	Indefinite
KQD-IC$^+$/ IKQD-IC$^+$	7.1 (1.4)	14.0 (2.8)
KQD-IC$^-$/ IKQD-IC$^-$	6.7 (0.8)	12.0 (2.1)
KQD-RC$^+$/ IKQD-RC$^+$	9.0 (2.1)	13.3 (3.4)
KQD-RC$^-$/ IKQD-RC$^-$	11.2 (2.2)	13.3 (4.1)
KQD-FK$^+$/ IKQD-FK$^+$	6.6 (0.6)	13.9 (2.7)
KQD-FK$^-$/ IKQD-FK$^-$	6.6 (1.1)	14.0 (2.2)
KFD / IKFD	10.5 (2.2)	12.5 (3.7)
SVM / ISVM	8.5 (1.5)	20.0 (3.2)
KPCA-QD / IKPCA-QD	8.4 (1.4)	12.6 (2.1)
KNN / IKNN	10.0 (2.1)	14.5 (1.9)

Numbers in parenthesis denote standard deviations.

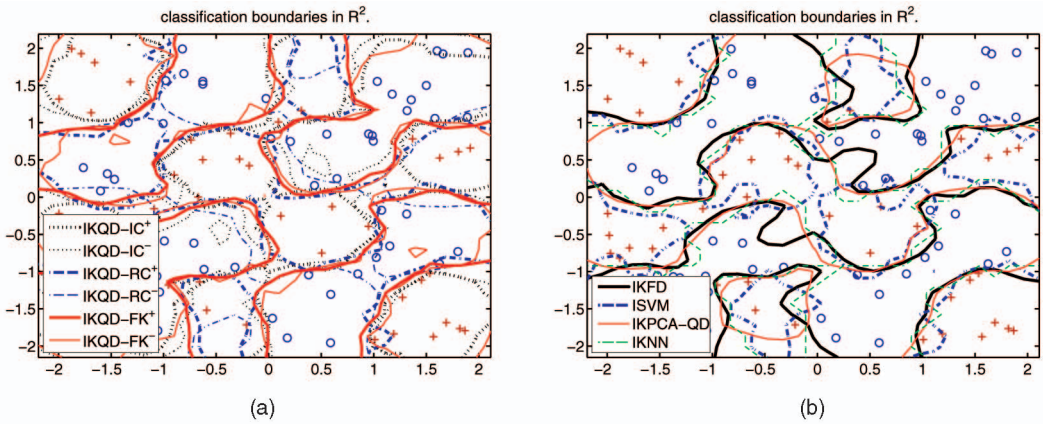


Fig. 2. Classifiers on 2D checkerboard data based on an invariant and indefinite Gaussian RBF-kernel. The classifier parameters are determined via cross-validation. (a) Results for all IKQD methods. (b) Results for the reference methods.

TABLE 5
Indices of Indefiniteness and Average Classification Errors [in Percent] for Different Kernel-Based Classifiers Based on the Invariant Gaussian RBF Kernel for Checkerboard Data The kernel parameter s Varies

Indefiniteness	Kernel parameter s						
	0.05	0.1	0.5	1	5	10	50
r_{neg}	0.162	0.181	0.213	0.222	0.217	0.210	0.130
(p, q)	(57, 43)	(54, 46)	(53,47)	(52,48)	(51,49)	(51,49)	(50,50)
$\|\|\phi_\mu^{[1]} - \phi_\mu^{[2]}\|\|^2_{\mathcal{K}}$	0.169	0.190	0.167	0.094	-0.054	-0.012	0.078
Classifier	0.05	0.1	0.5	1	5	10	50
IKQD-IC$^+$	49.0 (5.3)	41.8 (6.2)	14.6 (2.7)	15.2 (2.5)	20.4 (4.0)	23.1 (5.1)	21.2 (3.2)
IKQD-IC$^-$	48.8 (4.6)	36.6 (5.0)	14.3 (1.7)	12.5 (2.5)	14.7 (2.9)	18.4 (3.5)	18.4 (2.0)
IKQD-RC$^+$	15.1 (2.9)	13.3 (2.9)	19.4 (4.8)	33.4 (6.8)	47.8 (5.6)	51.1 (4.7)	35.8 (5.8)
IKQD-RC$^-$	14.6 (2.7)	14.5 (3.1)	14.2 (4.2)	16.5 (6.3)	47.0 (6.3)	51.0 (2.9)	36.0 (5.8)
IKQD-FK$^+$	14.0 (2.8)	14.3 (2.8)	12.2 (2.5)	12.3 (1.9)	13.8 (3.1)	13.1 (1.6)	18.2 (1.9)
IKQD-FK$^-$	18.9 (4.0)	16.9 (3.0)	14.8 (2.2)	12.7 (1.5)	14.3 (2.7)	12.9 (1.9)	17.8 (2.2)
IKFD	12.8 (1.7)	10.7 (2.9)	14.2 (2.4)	14.8 (4.1)	14.0 (2.4)	12.6 (2.2)	20.4 (1.3)
ISVM	20.2 (3.7)	21.0 (4.5)	26.5 (4.0)	46.6 (4.4)	44.7 (10.2)	49.3 (4.0)	31.7 (4.8)
IKNN	14.3 (2.1)	15.2 (2.4)	14.9 (2.5)	15.2 (2.5)	15.3 (2.0)	13.8 (2.3)	14.4 (2.3)
IKPCA-QD	14.2 (3.8)	15.0 (4.5)	12.2 (2.5)	12.1 (2.4)	14.0 (4.0)	12.2 (1.9)	17.8 (2.3)

Averaging is performed over 10 data drawings.

measured by the sums of absolute eigenvalues λ_i of K, and the squared distance of the class means $\|\phi_\mu^{[1]} - \phi_\mu^{[2]}\|^2_{\mathcal{K}}$.

Concerning indefiniteness, we note that the fraction of negative energy is the highest in the middle range of s and is decreasing toward both lower and higher values. This is expected because kernel matrices converge to either I_n for $s \to 0$ or to the matrix $\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ for $s \to \infty$, which are both positive semidefinite. Note that the square distance between class means in the embedded Krein space may by negative for some s. This gives rise to difficult separation with indefinite SVM [11]. Indeed, ISVM performs badly in these cases. We can observe that the IKQD-RC approaches seem to favor smaller values of s, whereas the IKQD-IC classifiers are better for larger s. The IKQD-FK approaches work acceptable over the whole range of s. KNN, despite of being a kernel classifier, is theoretically independent of the choice of s for the RBF-kernel. However, due to differing randomization seeds and numerical inaccuracies, the numbers in the table are slightly varying. Similar observations can be made for the positive definite data set of the previous section.

5.3 Real-World Kernel Data

We now consider both two-class and multiclass problems, ranging from positive definite kernels, slightly indefinite kernels to strongly indefinite kernels, and covering equally balanced as well as unbalanced class sizes. We compare the performance of the IKQD methods to the reference classifiers.

The data are defined either by a symmetric dissimilarity function $d(x, x')$ or symmetric similarity function $s(x, x')$, designed or optimized for the given task. Examples of such measures are edit distance, variants of Hausdorff distances, compression distance, structural similarity, or shape matching similarity. These pairwise functions allows us to define suitable kernels by $k(x, x') := -(d(x, x'))^2$ or $k(x, x') := s(x, x')$ after appropriate linear scaling. The scaling is done such that all dissimilarities are divided by the average dissimilarity in the training set, or by the average self-similarity if we deal with similarity data. Such a scaling is only important for practical reasons in order to use identical ranges of cross-validated parameters for different data sets.

The centered training kernel matrix $\tilde{K}$ obtained from a dissimilarity function is positive definite *only if* the dissimilarity matrix $D := (d(x_i, x_j))_{i,j=1}^n$ is isometrically embedabble into a euclidean space [10], [26]. Since this does not often occur for optimized proximities, we will mostly encounter indefinite kernels. Consequently, we use the indefinite notation throughout this section for all the IKQD techniques and reference classifiers.

TABLE 6
Characteristics of Real-World Kernel Data β Is the Fraction of Data Used for Training in the Holdout Experiments

	Dissimilarity	Kernel	c (n_j)	β	$r_{\text{neg}}(p,q)$
Two-class problems					
Mucosa	Derivative l_1	$-d^2$	2 (132/856)	0.60	0.15 (216,378)
Heart	Euclidean	$-d^2$	2 (139/164)	0.80	0.00 (242, 0)
Nist38-EU	Euclidean	$-d^2$	2 (1000)	0.10	0.00 (199, 0)
Nist38-MH	Mod. Hausd.	$-d^2$	2 (1000)	0.10	0.22 (104, 95)
Poly-H	Hausdorff	$-d^2$	2 (2000)	0.05	0.32 (113, 87)
Poly-MH	Mod. Hausd.	$-d^2$	2 (2000)	0.05	0.25 (91,108)
Multi-class problems					
Cat-cortex	Prior knowl.	$-d^2$	4 (10–19)	0.80	0.19 (35, 18)
Protein	Evolutionary	$-d^2$	4 (30–77)	0.80	0.00 (167, 3)
News-COR	Correlation	$-d^2$	4 (102–203)	0.60	0.19 (127,208)
ProDom	Structural	s	4 (271–1051)	0.25	0.01 (518, 90)
Chicken15	Edit-dist.	$-d^2$	5 (61–117)	0.80	0.27 (202,156)
Chicken29	Edit-dist.	$-d^2$	5 (61–117)	0.80	0.31 (192,166)
Files	Compression	$-d^2$	5 (60–255)	0.50	0.02 (392, 63)
Pen-ANG	Edit-dist.	$-d^2$	10 (334–363)	0.15	0.24 (261,269)
Pen-DIS	Edit-dist.	$-d^2$	10 (334–363)	0.15	0.28 (253,276)
Zongker	Shape-match.	s	10 (200)	0.25	0.36 (274,226)
Chromo-DIF	Edit-dist.	$-d^2$	21 (200)	0.10	0.21 (206,213)
Chromo-ABS	Edit-dist.	$-d^2$	21 (200)	0.10	0.18 (198,221)

The indices of indefiniteness of the kernel, $r_{\text{neg}} \in [0,1]$ and (p,q), are averaged over 25 runs.

The data sets are described in Appendix B, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.290, while kernel matrices are briefly characterized in Table 4. Note that the indefiniteness indices r_{neg} and (p,q) are derived on the centered kernels (as the IKQD and IKFD methods rely on either global or classwise centering).

We run holdout experiments in which the complete data set is split into the training and test kernel matrices such that the specified β-fraction is used for training (see Table 6). We do not set a fixed β because the data sets have variable sizes while we want to focus on small-size and moderate-size problems (due to time complexity as well). In each run, parameters of all classifiers are determined by 10-fold cross-validation. Here, the classwise regularization parameters are again kept identical for all classes, i.e., $\alpha_j := \alpha$ and $\sigma_j^2 := \sigma^2$. The following parameter ranges are considered: $\alpha \in [10^{-6}, 0.5]$ for IKFD, IKQD-IC$^+$, and IKQD-FK$^+$, $\alpha \in [10^{-8}, 0.5]$ for IKPCA-QD, IKQD-IC$^-$, and IKQD$-$FK$^-$, $\sigma^2 \in [10^{-6}, 2]$ for the IKQD-RC approaches, and $C \in [10^{-1}, 10^8]$ for ISVM. The total number of investigated values is 11-13. For IKNN, the value $k \in \{1, 2, \ldots, 45\}$ is optimized. IKPCA-QD has two parameters: the amount of preserved variance p_{var} in the IKPCA and the regularization α of QDA in the IKPCA space. We set $p_{\text{var}} = 0.8$ in all experiments as IKPCA usually gives very long eigenvalue tails which are not very informative. The

TABLE 7
Average Classification Errors (%) for Positive-Definite and Indefinite Kernel Data Numbers in Parentheses Denote Standard Deviations

Two-class problems						
	Mucosa	Heart	Nist38-EU	Nist38-MH	Poly-H	Poly-MH
IKQD-IC$^+$	21.0 (2.2)	50.0 (0.0)	3.8 (0.8)	10.7 (1.5)	20.5 (2.0)	19.8 (2.1)
IKQD-IC$^-$	21.0 (2.3)	50.0 (0.0)	4.5 (1.0)	12.0 (1.3)	24.3 (3.0)	21.3 (2.8)
IKQD-RC$^+$	27.6 (6.3)	17.9 (3.0)	6.7 (0.9)	10.7 (1.3)	**5.4 (1.3)**	2.7 (0.7)
IKQD-RC$^-$	45.0 (6.5)	**17.5 (4.4)**	7.6 (1.0)	12.4 (1.6)	6.6 (1.9)	2.8 (0.9)
IKQD-FK$^+$	**12.6 (2.2)**	25.6 (4.4)	**3.8 (0.7)**	**5.9 (1.3)**	6.6 (1.2)	**1.5 (0.4)**
IKQD-FK$^-$	19.3 (4.1)	50.0 (0.0)	4.5 (1.1)	7.7 (1.2)	10.7 (2.0)	2.5 (0.5)
IKFD	18.7 (1.4)	**15.4 (3.4)**	**4.0 (0.7)**	7.5 (1.4)	**6.5 (1.0)**	**0.9 (0.4)**
ISVM	**9.0 (1.1)**	19.5 (4.5)	7.7 (0.5)	15.6 (0.9)	21.6 (7.7)	7.6 (2.4)
IKNN	22.3 (2.9)	17.7 (3.6)	6.4 (0.6)	**6.4 (0.8)**	7.0 (1.5)	5.6 (0.9)
IKPCA-QD	23.1 (4.1)	20.4 (4.4)	7.2 (0.4)	6.9 (1.0)	7.2 (1.6)	2.2 (0.4)
Multi-class problems						
	Cat-cortex	Protein	News-COR	Prodom	Chicken15	Chicken29
IKQD-IC$^+$	84.3 (12.7)	34.3 (8.6)	74.5 (4.4)	70.0 (3.8)	37.7 (4.1)	30.8 (3.8)
IKQD-IC$^-$	84.2 (13.7)	35.5 (8.2)	73.6 (2.8)	70.5 (4.4)	40.2 (5.1)	34.0 (3.7)
IKQD-RC$^+$	8.7 (9.1)	1.5 (2.8)	**24.1 (2.4)**	1.5 (0.7)	**7.0 (2.8)**	**5.3 (2.4)**
IKQD-RC$^-$	**7.0 (7.1)**	1.3 (2.8)	24.2 (2.6)	**1.5 (0.6)**	14.3 (4.9)	6.1 (2.4)
IKQD-FK$^+$	7.3 (7.4)	**0.4 (1.7)**	26.1 (2.8)	2.0 (1.0)	15.7 (3.7)	9.3 (1.9)
IKQD-FK$^-$	27.8 (14.8)	1.6 (2.3)	45.7 (4.4)	4.3 (2.2)	28.1 (3.8)	25.5 (3.9)
IKFD	13.0 (10.3)	0.6 (2.5)	25.8 (2.6)	1.8 (0.6)	11.3 (2.9)	12.9 (2.5)
ISVM	32.0 (9.5)	8.5 (9.7)	**23.3 (2.5)**	6.9 (10.9)	22.9 (3.8)	16.0 (3.3)
IKNN	16.3 (9.9)	3.6 (3.5)	29.7 (2.7)	3.0 (0.6)	**8.5 (2.9)**	**4.7 (2.7)**
IKPCA-QD	**10.5 (10.6)**	**0.5 (1.1)**	26.5 (2.7)	**1.3 (0.5)**	17.9 (3.7)	14.0 (2.4)
	Files	Pen-ANG	Pen-DIST	Zongker	Chromo-DIF	Chromo-ABS
IKQD-IC$^+$	64.9 (10.6)	4.3 (0.6)	5.4 (1.2)	39.7 (1.6)	43.4 (4.2)	26.4 (3.6)
IKQD-IC$^-$	64.0 (9.4)	5.2 (0.7)	6.4 (1.1)	45.7 (2.6)	60.5 (3.7)	47.0 (5.5)
IKQD-RC$^+$	**6.2 (1.8)**	6.6 (1.2)	11.8 (2.1)	5.6 (0.7)	**6.0 (0.8)**	**9.1 (1.0)**
IKQD-RC$^-$	6.8 (2.0)	11.0 (2.5)	18.2 (2.4)	5.6 (0.9)	6.4 (1.1)	10.8 (1.1)
IKQD-FK$^+$	6.9 (1.6)	**3.3 (1.0)**	**3.0 (0.9)**	**4.4 (0.6)**	40.7 (6.5)	30.8 (6.0)
IKQD-FK$^-$	17.3 (4.0)	3.9 (1.0)	3.5 (1.0)	31.4 (4.2)	81.4 (3.1)	81.0 (3.3)
IKFD	**6.6 (1.4)**	1.4 (0.5)	1.5 (0.5)	**5.8 (0.6)**	8.6 (0.8)	**7.7 (0.4)**
ISVM	8.9 (2.2)	41.0 (2.5)	42.0 (2.2)	92.9 (1.5)	89.0 (1.6)	87.1 (2.2)
IKNN	36.3 (3.3)	1.1 (0.5)	1.7 (0.5)	11.5 (1.4)	**7.7 (0.5)**	8.0 (0.7)
IKPCA-QD	14.1 (2.5)	**1.1 (0.4)**	**1.4 (0.3)**	6.6 (0.7)	8.6 (0.7)	9.7 (0.8)

Best IKQD and reference classifier are highlighted in each column.

complete procedure is repeated 25 times for all classifiers and the results are averaged out.

Table 7 shows average classification errors and standard-deviations for the IKQD classifiers and reference methods. Problems with nearly pd kernels are: *Nist38-EU* (pd), *Heart* (pd), *Protein*, *Prodom*, and *Files*. Problems with moderately indefinite kernels are: *Mucosa*, *Chromo-ABS*, *Cat-cortex*, *News-COR*, *Chromo-DIF*, and *Nist38-MH*, while the remaining problems deal with highly indefinite kernels. The following observations can be made for the IKQD methods:

- All IKQD-$*^+$ approaches perform usually similarly or better than their corresponding IKQD-$*^-$ variants.
- IKQD-IC$^+$ and IKQD-IC$^-$ frequently perform badly.
- Usually, one of the best methods is either IKQD-RC$^+$ or IKQD-FK$^+$.

Among the reference methods, we see that:

- ISVM performs badly for multiclass indefinite kernel problems and is mostly outperformed by IKFD or IKNN.
- IKFD works, in general, very well with indefinite kernels, which is an empirical support in addition to its sound geometrical motivation.
- There is no clear favorite among the reference classifiers IKFD, ISVM, IKNN, and IKPCA-QD.

By comparing our IKQD approaches and the reference classifiers, we conclude that:

- Overall, in half of the cases, a reference method gives better performance than all the IKQD methods.
- ISVM is outperformed by IKQD-FK$^+$ in all cases except for the *Heart* and *Mucosa* data.
- IKNN is outperformed by IKQD-FK$^+$ in all but the *Chicken-*$*$, *Pen-*$*$ and *Chromo-*$*$ examples.
- IKQD-FK$^+$ outperforms IKPCA-QD for a small number of classes c. IKPCA-QD tends to work better than IKQD-FK$^+$ if $c \geq 10$.
- IKFD achieves better results than IKQD-RC$^+$ in 9 out of the 18 data sets and outperforms IKQD-FK$^+$ in 8 cases.
- ISVM is usually significantly outperformed by either IKQD-RC$^+$ or IKQD-FK$^+$.

These findings are also supported by further experiments on positive definite kernels, resulting from vectorial data with Gaussian RBF kernel, which we omit here.

6 Summary and Conclusions

In this paper, we have presented different formulations for kernel quadratic discriminants. In particular, we make a distinction between approaches based on invertible covariance operators KQD-IC, regularized covariance operators KQD-RC, and full kernel space approaches KQD-FK. All methods rely on kernel Mahalanobis distances, appropriately regularized in kernel-induced feature spaces. They differ in the amount of kernel information they rely on. When ignoring the computation of the bias b_j, the KQD-IC and KQD-RC approaches do not use the between-class kernel submatrices. As a result, lower than expected recognition accuracy may be achieved as the methods can only work well for the classes with "clean" separation (as, e.g., in the *Nist38-EU* case). This means that the between-class kernel values are much smaller than the within-class kernel values. In contrast, the KQD-FK methods rely on the full kernel information for the computation of the Mahalanobis distance. In addition to the test-versus-train matrix, the KQD-RC approaches require the diagonal kernel values $k(x,x)$. A formal limitation of KQD-IC$^-$ and IKQD-IC$^-$ is that the assumption of invertible covariance operator is made for the derivation, though not required, and hence, not checked for the final classifier evaluation. Still a failure of this assumption on certain data sets may lead to a loss of recognition accuracy.

Concerning computation complexities, the KQD-IC and KQD-RC approaches have the conceptual advantage of a reduced test time for large number of classes. The dominating complexity contributions are inversely growing with the number of classes. However, except for KQD-RC$^-$, the classification time of the methods grows quadratically with n in contrast to linear kernel methods. Future work will aim at acceleration, e.g., by sparse matrix approximations for the inverses of covariance matrices or training subset selection. KQD is a true multiclass approach, not depending on series of binary decisions. As the computation schemes are identical for all classes, the decisions functions can easily be parallelized.

The methods are genuinely nonlinear, which is conceptually wider and may be favorable in comparison to kernel methods obtained from linear algorithms. The methods have natural extensions to indefinite kernels. In particular, we present a derivation of indefinite KFD, which has a geometric interpretation in Krein spaces. We also propose extensions of all discussed KQD discriminants to indefinite kernels. All these methods have a sound mathematical motivation; hence, they extend the class of kernel methods to the methods that work with general, both positive and indefinite, kernels.

Experimentally, IKQD-RC$^+$ and IKQD-FK$^+$ seem to be favorable among the IKQD approaches. The latter method seems the most beneficial, but is computationally more expensive due to the processing of full kernel matrix for each discriminant. The IKQD-IC$^+$ and IKQD-IC$^-$ techniques do not perform well in some cases. In addition to the conceptional arguments given earlier, there may be numerical difficulties caused by the use of the second power of (pseudo)-inverses of the class-related diagonal kernel submatrices (see (17) and (16)). If there is insufficient discriminative information in the class-related kernel submatrix, it will be enhanced in this process. IKFD is frequently similar or better than the IKQD approaches, but there are also many situations in which IKQD-RC$^+$ or IKQD-FK$^+$ are strong winners. This especially occurs for suboptimally designed dissimilarity measures, as encountered in the *Nist38-MH*, *Poly-H*, *Chicken-15*, and *Chromo-DIF* cases, or for imbalanced data such as the *Cat-cortex* or *News-COR* cases. In general, the IKQD-RC and IKQD-FK methods mostly outperform ISVM, which becomes apparent with a growing indefiniteness of the kernel. The best IKQD method, IKQD-FK$^+$, frequently outperforms the reference classifiers IKNN and IKPCA-QD.

In summary, we provide a comprehensive approach to kernel quadratic discriminant analysis based on the suitably regularized kernel Mahalanobis distance in either class-related or full kernel-induced subspaces. The bias terms in the derived discriminant functions are currently found on the training set such that they minimize the training error. This is done in order to avoid numerical problems that arise if we want to express them in analogy to the quadratic discriminant in a Euclidean space, i.e., as the logarithms of the eigenvalue-product of the scaled kernel matrices. There is also some room left for a possibly better and/or more reliable estimation of the bias terms. More research is also needed to clearly identify conditions under which the nonlinear KQD/IKQD methods will outperform the linear KFD/IKFD and SVM/ISVM techniques.

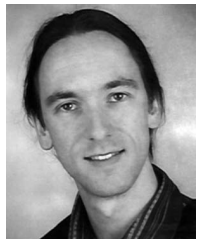
ACKNOWLEDGMENTS

This work is supported by the Engineering and Physical Science Research Council in the UK, project no. EP/D066883/1 and the German Academic Exchange Service (DAAD), contract no. D/07/09940.

REFERENCES

[1] J. Bognár, *Indefinite Inner Product Spaces.* Springer-Verlag, 1974.
[2] S. Canu, X. Mary, and A. Rakotomamonjy, "Functional Learning Through Kernel," *Advances in Learning Theory: Methods, Models and Applications,* J. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, eds., pp. 89-110, IOS Press, 2003.
[3] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines,* http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
[4] R. Der and D. Lee, "Large-Margin Classification in Banach Spaces," *Proc. Int'l Conf. Artificial Intelligence and Statistics,* vol. 2, pp. 91-98, 2007.
[5] M. Dritschel and J. Rovnyak, "Operators on Indefinite Inner Product Spaces," *Lectures on Operator Theory and Its Applications, Fields Inst. Monographs,* pp. 141-232, 1996.
[6] M. Dubuisson and A. Jain, "Modified Hausdorff Distance for Object Matching," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 566-568, 1994.
[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification,* second ed. John Wiley & Sons, Inc., 2001.
[8] R. Duin, P. Juszczak, D. de Ridder, P. Paclík, E. Pekalska, and D. Tax, "PR-Tools," http://prtools.org, 2004.
[9] L. Goldfarb, "A New Approach to Pattern Recognition," *Progress in Pattern Recognition,* L. Kanal and A. Rosenfeld, eds., vol. 2, pp. 241-402, Elsevier Science Publishers, 1985.
[10] J. Gower, "Metric and Euclidean Properties of Dissimilarity Coefficients," *J. Classification,* vol. 3, pp. 5-48, 1986.
[11] B. Haasdonk, "Feature Space Interpretation of SVMs with Indefinite Kernels," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 5, pp. 482-492, May 2005.
[12] B. Haasdonk, "Transformation Knowledge in Pattern Analysis with Kernel Methods-Distance and Integration Kernels," PhD dissertation, Universität Freiburg, Institut für Informatik, 2005.
[13] B. Haasdonk and H. Burkhardt, "Invariant Kernel Functions for Pattern Analysis and Machine Learning," *Machine Learning,* vol. 68, no. 1, pp. 35-61, 2007.
[14] B. Haasdonk and E. Pekalska, "Indefinite Kernel Fisher Discriminant," *Proc. Int'l Conf. Pattern Recognition,* 2008.
[15] B. Hassibi, A. Sayed, and T. Kailath, "Linear Estimation in Krein Spaces—Part I: Theory," *IEEE Trans. Automatic Control,* vol. 41, no. 1, pp. 18-33, 1996.
[16] M. Hein, O. Bousquet, and B. Schölkopf, "Maximal Margin Classification for Metric Spaces," *J. Computer and System Sciences,* vol. 71, no. 3, pp. 333-359, 2005.
[17] S. Hettich, C. Blake, and C. Merz, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.
[18] S. Hochreiter and K. Obermayer, "Support Vector Machines for Dyadic Data," *Neural Computation,* vol. 18, no. 6, pp. 1472-1510, 2006.
[19] S.-Y. Huang, C.-R. Hwang, and M.-H. Lin, "Kernel Fisher's Discriminant Analysis in Gaussian Reproducing Kernel Hilbert Space," technical report, Academia Sinica, Taipei, Taiwan, 2005.
[20] D. Jacobs, D. Weinshall, and Y. Gdalyahu, "Classification with Non-Metric Distances: Image Retrieval and Class Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 6, pp. 583-600, June 2000.
[21] J. Laub and K.-R. Müller, "Feature Discovery in Non-Metric Pairwise Data," *J. Machine Learning Research,* pp. 801-818, 2004.
[22] X. Mary, "Moore-Penrose Inverse in Krein Spaces," *Integral Equations and Operator Theory,* pp. 419-433, 2008.
[23] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher Discriminant Analysis with Kernels," *Proc. Neural Networks for Signal Processing,* pp. 41-48, 1999.
[24] S. Mika, A. Smola, and B. Schölkopf, "An Improved Training Algorithm for Kernel Fisher Discriminants," *Proc. Int'l Conf. Artificial Intelligence and Statistics,* pp. 98-104, 2001.
[25] C. Ong, X. Mary, S. Canu, and S.A.J. Smola, "Learning with Non-Positive Kernels," *Proc. Int'l Conf. Machine Learning,* pp. 639-646, 2004.
[26] E. Pekalska and R. Duin, *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications.* World Scientific, 2005.
[27] E. Pekalska and R. Duin, "Indefinite Kernel PCA," work in progress, 2009.
[28] E. Pekalska, A. Harol, R. Duin, B. Spillmann, and H. Bunke, "Non-Euclidean or Non-Metric Measures Can be Informative," *Proc. Joint IAPR Workshops Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition,* pp. 871-880, 2006.
[29] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning.* MIT Press, 2006.
[30] V. Roth, J. Laub, J. Buhmann, and K.-R. Müller, "Going Metric: Denoising Pairwise Data," *Proc. Advances in Neural Information Processing Systems,* pp. 841-856, 2003.
[31] J. Rovnyak, "Methods of Krein Space Operator Theory," *Operator Theory: Advances and Applications,* vol. 134, pp. 31-66, 2002.
[32] A. Ruiz and P. Lopez-de Teruel, "Nonlinear Kernel-Based Statistical Pattern Analysis," *IEEE Trans. Neural Networks,* vol. 12, no. 1, pp. 16-32, 2001.
[33] B. Schölkopf and A. Smola, *Learning with Kernels.* MIT Press, 2002.
[34] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, no. 5, 1998.
[35] B. Schölkopf, K. Tsuda, and J. Vert, *Kernel Methods in Computational Biology.* MIT Press, 2004.
[36] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge Univ. Press, 2004.
[37] P. Simard, Y.A. Le Cun, J.S. Denker, and B. Victorri, "Transformation Invariance in Pattern Recognition—Tangent Distance and Tangent Propagation," *Int'l J. Imaging System and Technology,* vol. 11, no. 3, pp. 181-194, 2001.
[38] V. Vapnik, *Statistical Learning Theory.* John Wiley & Sons, Inc., 1998.
[39] U. von Luxburg and O. Bousquet, "Distance-Based Classification with Lipschitz Functions," *J. Machine Learning Research,* vol. 5, pp. 669-695, 2004.
[40] G. Wahba, "Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV," *Advances in Kernel Methods, Support Vector Learning,* B. Schölkopf, C. Burges, and A. Smola, eds., pp. 69-88, MIT Press, 1999.
[41] J. Wang, K. Plataniotis, J. Lu, and A. Venetsanopoulos, "Kernel Quadratic Discriminant Analysis for Small Sample Size Problem," *Pattern Recognition,* vol. 41, no. 5, pp. 1528-1538, 2008.

Elżbieta Pekalska received the MSc degree in computer science from the University of Wroclaw, Poland, in 1996 and the PhD degree (cum laude) in computer science from Delft University of Technology, Delft, The Netherlands, in 2005. During 1998-2004, she was with Delft University of Technology, The Netherlands, where she worked on both fundamental and applied projects in pattern recognition. She is currently an engineering and physical sciences research council fellow at the University of Manchester, United Kingdom. She is engaged in learning processes and learning strategies, as well as in the integration of bottom-up and top-down approaches, which not only includes intelligent learning from data and sensors, but also human learning on their personal development paths. She is the author or coauthor of more than 40 publications, including a book, journal articles, and international conference papers. Her current research interests focus on the issues of representation, generalization, combining paradigms, and the use of kernels and proximity in the learning from examples. She is also involved in the understanding of brain research, neuroscience, and psychology.



Bernard Haasdonk studied physics, mathematics, and computer science at the University of Freiburg in the years 1995-2001. After completing his master's thesis in numerical analysis in 2001, he started research in pattern recognition and machine learning. One particular focus of his work represents kernel methods and kernel design. Since his PhD thesis in 2005, the continuation of this research has been partially supported by a scholarship from the German Scientific Exchange Service (DAAD). He extended his focus to the field of model reduction of numerical simulation methods and joined the Applied Mathematics Institute at the University of Freiburg as a postdoctoral researcher. He spent some months at the Massachusetts Institute of Technology and moved to the University of Münster in 2007. This recent research was supported by a scholarship of the "Landesstiftung Baden Württemberg gGmbH." In 2009, he joined the "Excellence Cluster Simulation Technology" at the University of Stuttgart as a junior professor of mathematics.