ANGEWANDTE MATHEMATIK

UND

INFORMATIK

Kernel Quadratic Discriminant Analysis for Positive Definite and Indefinite Kernels

Elżbieta Pękalska and Bernard Haasdonk

06/08 - N

UNIVERSITÄT MÜNSTER

Kernel Quadratic Discriminant Analysis for Positive Definite and Indefinite Kernels

Elżbieta Pękalska and Bernard Haasdonk

***Abstract*— Kernel methods are well established and successful algorithms for pattern analysis thanks to their mathematical elegance and efficient solutions they provide. Numerous nonlinear extensions of pattern recognition techniques have been proposed based on the so-called kernel trick.**

The objective of this paper is twofold. First, we derive an additional kernel tool, namely kernel quadratic discriminant (KQD). We discuss different formulations of KQD based on the regularized kernel Mahalanobis distance in both full and class-related subspaces. Secondly, 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 disobey 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**

I. Introduction

Kernel methods are powerful statistical learning techniques [33], [31], widely applied to various learning scenarios thanks 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 [28], [34]. Thanks 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 in their kernel-based formulations. These include Support Vector Machines (SVM), kernel PCA, kernel Fisher discriminant (KFD), kernel k-means, etc. [31]. 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 non-vectorial data descriptions, such as strings, bags of words, graphs, shapes, probability models [30], [31], 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. This may occur when invariance or robustness is incorporated into the measure [32], [16], [12], due to suboptimal optimization procedures for measure derivation [23], partial projections or occlusions [16], context-dependent alignments or object comparisons [5], [25], or derived from intrinsic non-Euclidean or non-metric dissimilarities, such as modified Hausdorff distances [5] or non-pd similarities, such as Kullback-Leibler divergence. Consequently, there is a practical need to properly handle these measures. While many researchers choose to regularize non-pd kernels to make them pd, a natural extension of Mercer kernels leads to indefinite (Kreĭn) kernels [2], [20], [17], [10], [21] or dyadic kernels [14]. Both these are examples of proximity representations, i.e. matrices whose elements encode degrees of similarity between pairs of examples or examples and optimized prototypes [21]. Therefore, developing and investigating methods for indefinite kernels is of high interest. In particular, a further contribution in this paper lies in the extension of kernel linear and quadratic discriminants to indefinite kernels. We provide a sound underpinning of the methods. Experiments on toy and real-world data show the good performance of the KQD method for both positive definite and indefinite kernels.

The paper is organized as follows. Section II starts with preliminaries on kernels. Section III presents the indefinite kernel Fisher discriminant analysis. Section IV is the main part that introduces different formulations of KQD analysis for both positive definite and indefinite kernels. Section V focuses on an experimental study illustrating performance of kernel discriminant analysis on toy and real world data. The final discussion is presented in Section VI. To maintain clarity, the detailed derivations of the methods are left out from the main text, but included in Appendix I.

II. Preliminaries on kernels

We will first introduce some notation and provide basics for Hilbert spaces and positive definite kernels. Then we will focus on Kreĭn spaces and indefinite kernels.

A. *Positive definite kernels*

Assume $\mathcal{X}$ 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 $\xi_1^{\mathsf{T}} \xi_2 := \langle \xi_1, \xi_2 \rangle_{\mathcal{H}}$. 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, $\xi \mathbf{v}^{\mathsf{T}} = [v_1 \xi, \ldots, v_m \xi]$ for a single $\xi \in \mathcal{H}$ and vector $\mathbf{v}$. For two sequences $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_m], \boldsymbol{\xi}' = [\xi_1', \ldots, \xi_n']$ in $\mathcal{H}$, we

E. Pękalska is with the School of Computer Science, University of Manchester, United Kingdom; e-mail: pekalska@cs.man.ac.uk.

B. Haasdonk is with the Institute of Numerical and Applied Mathematics, University of Münster, Germany; e-mail: haasdonk@math.uni-muenster.de.

write $G := \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi}' \in \mathbb{R}^{m\times n}$ for a cross-Gram matrix with entries $G_{ij} = \langle \xi_i, \xi'_j \rangle_{\mathcal{H}}$.

We address a c-class problem, given by the training data $X_{\mathrm{tr}} = \{x_i\}_{i=1}^n \subset \mathcal{X}$ with labels $\{y_i\}_{i=1}^n \subset \Omega$, where $\Omega := \{\omega_1, \dots, \omega_c\}$ is a set of c target classes. Let $\Phi := [\phi(x_1), \dots, \phi(x_n)]$ be a sequence of images of the training data X_{tr} in $\mathcal{H}$. Without loss of generality the vectors in Φ are grouped into classes such that $\Phi = [\Phi^{[1]}, \Phi^{[2]}, \dots, \Phi^{[c]}]$, where $\Phi^{[j]} := [\phi(x_1^j), \dots, \phi(x_{n_j}^j)]$ represents the n_j-element class ω_j and $\sum_{j=1}^c n_j = n$.

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

$$C = \frac{1}{n}\,\tilde{\Phi}\tilde{\Phi}^\mathsf{T} = \frac{1}{n}\Phi H H \Phi^\mathsf{T}.$$

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

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

The transformation ϕ acts as a (usually) non-linear 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 $k(x, x') = \phi(x)^\mathsf{T}\phi(x')$ for any $x, x' \in \mathcal{X}$. Consequently, $K := \Phi^\mathsf{T}\Phi$ is an $n \times n$ kernel matrix derived from the training data. Moreover, we can also define the centered kernel matrix $\tilde{K} := \tilde{\Phi}^\mathsf{T}\tilde{\Phi} = H\Phi^\mathsf{T}\Phi H = HKH$. In addition to the quantities defined for the complete training sequence Φ, we can define analogous class-wise quantities for $\Phi^{[j]}, j = 1, \dots, c$, which are consequently indicated with the superscript $[j]$. Further, for an arbitrary $x \in \mathcal{X}$, $\mathbf{k}_x$ denotes the vector of kernel values of x to the training data, while $\tilde{\mathbf{k}}_x$ is the centered vector:

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

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

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

B. Indefinite kernels

The above terminology and notation can be extended to Kreĭn spaces; see [1], [8], [4], [26], [21] for details. 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. $\langle \xi_+, \xi_- \rangle_{\mathcal{K}} = 0$ holds 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 non-zero 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 *fundamental projections*. Any $\xi \in \mathcal{K}$ can be represented as $\xi = P_+\,\xi + P_-\,\xi$ and $I_{\mathcal{K}} = P_+ + P_-$ is the identity operator. The linear operator $\mathcal{J} = P_+ - P_-$ is called the *fundamental symmetry* and is the basic characteristic of a Kreĭn space $\mathcal{K}$, satisfying $\mathcal{J} = \mathcal{J}^{-1} = \mathcal{J}^\mathsf{T}$. The space $\mathcal{K}$ can be turned into its *associated Hilbert space* $|\mathcal{K}|$ by using the positive definite inner product $\langle \xi, \xi' \rangle_{|\mathcal{K}|} := \langle \xi, \mathcal{J}\xi' \rangle_{\mathcal{K}}$. A countable orthonormal basis $\mathcal{B}_+$ for $\mathcal{K}_+$ and $\mathcal{B}_-$ for $\mathcal{K}_-$ yields a basis $\mathcal{B} := \mathcal{B}_+ \cup \mathcal{B}_-$ for $\mathcal{K}$, which 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 pd case, we use the "transposition" abbreviation $\xi^\mathsf{T}\xi' := \langle \xi, \xi' \rangle_{|\mathcal{K}|}$ and now additionally (motivated by $\mathcal{J}$ operating as a sort of "conjugation") a "conjugate-transposition" notation $\xi^*\xi' := \langle \xi, \xi' \rangle_{\mathcal{K}} = \langle \mathcal{J}\xi, \xi' \rangle_{|\mathcal{K}|} = (\mathcal{J}\xi)^\mathsf{T}\xi' = \xi^\mathsf{T}\mathcal{J}\xi'$.

Finite-dimensional Kreĭn spaces with $\mathcal{K}_+ = \mathbb{R}^p$ and $\mathcal{K}^- = \mathbb{R}^q$ are denoted with $\mathbb{R}^{(p,q)}$ 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 $\mathbb{R}^{(p,q)}$. Kreĭn spaces are important as they provide feature-space representations of dissimilarity data [8] or indefinite kernels. For indefinite kernels, i.e. symmetric functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and finite data $\mathcal{X}$, the resulting kernel matrix K yields an embedding $\psi : \mathcal{X} \to \mathcal{K}$ into a finite-dimensional Kreĭn 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 Kreĭn space $\mathcal{K}$ and $\Psi := [\psi(x_1), \dots, \psi(x_n)]$ 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 Sec. II-A 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 $\tilde{\Psi} := [\tilde{\psi}(x_1), \dots, \tilde{\psi}(x_n)] = \Psi - \frac{1}{n}\Psi\mathbf{1}_n\mathbf{1}_n^\mathsf{T} = \Psi H$. 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

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

where $C^{|\mathcal{K}|} = \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^\mathsf{T}$ is the empirical covariance operator in $|\mathcal{K}|$.

TABLE I

KERNEL-INDUCED QUANTITIES FOR PD AND INDEFINITE KERNELS. DATA EMBEDDINGS Φ AND Ψ ARE IN KERNEL-INDUCED HILBERT AND KREĬN SPACES, RESPECTIVELY.

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

The operator C is not pd in the Hilbert sense, but in the Kreĭn sense [1], [26], i.e. $\langle \xi, C\xi \rangle_{\mathcal{K}} \geq 0$ for $\xi \neq 0$, hence in agreement with the inner product of this space. Assuming that C is invertible (which requires $n > \dim(\mathcal{K})$), the empirical square Mahalanobis distance $D_M^2(\cdot; \{\psi_\mu, C\}) : \mathcal{K} \to \mathbb{R}_{\geq 0}$ of a vector $\psi(x) \in \mathcal{K}$ to the data described by $\{\psi_\mu, C\}$ 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 $K := \Psi^* \Psi = \Psi^\top \mathcal{J} \Psi$. Similarly to traditional kernels, the centered kernel matrix is $\tilde{K} := \tilde{\Psi}^* \tilde{\Psi} = \tilde{\Psi}^\top \mathcal{J} \tilde{\Psi} = HKH$. 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 I summarizes these definitions for both types of kernels. In particular, $\mathcal{J} = I$ in the positive definite case, hence $\xi^* = \xi^\top$ and all definitions from this section reduce to the ones from Sec. II-A. Note that we could have focussed on the mere indefinite notation as the pd case is just 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.

III. KERNEL FISHER DISCRIMINANT ANALYSIS

Kernel Fisher discriminant (KFD) was proposed and successfully applied by Mika et al. [18], [19]. Since it is well known, we will directly focus on the extension to the indefinite case.

A. Indefinite Kernel Fisher Discriminant

Assume the training data for a two-class problem, $c = 2$, is embedded into a Kreĭn 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. Analogously 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

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

where the between-class scatter operator acts as $\Sigma_B^{\mathcal{K}} w = (\psi_\mu^{[1]} - \psi_\mu^{[2]}) \langle \psi_\mu^{[1]} - \psi_\mu^{[2]}, w \rangle_{\mathcal{K}} = (\psi_\mu^{[1]} - \psi_\mu^{[2]})(\psi_\mu^{[1]} - \psi_\mu^{[2]})^\top \mathcal{J} w$. Hence, $\Sigma_B^{\mathcal{K}} = \Sigma_B^{|\mathcal{K}|} \mathcal{J}$, where $\Sigma_B^{|\mathcal{K}|} = (\psi_\mu^{[1]} - \psi_\mu^{[2]})(\psi_\mu^{[1]} - \psi_\mu^{[2]})^\top$ 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 $\Sigma_W^{|\mathcal{K}|} := \sum_{j=1}^2 P(\omega_j) \sum_i (\psi(x_i^j) - \psi_\mu^{[j]})(\psi(x_i^j) - \psi_\mu^{[j]})^\top$ 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 to

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

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 $J(w) = v^\top \Sigma_B^{|\mathcal{K}|} v / (v^\top \Sigma_W^{|\mathcal{K}|} v)$ and $f(x) = v^\top \psi(x) + b$ with $b = -\frac{1}{2} v^\top (\psi_\mu^{[1]} + \psi_\mu^{[2]})$. This means that the Fisher discriminant in the Kreĭn 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.

A kernel method should avoid such explicit embeddings into a Kreĭn 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ encodes the inner product $k(x_i, x_j) = \psi(x_i)^\top \mathcal{J} \psi(x_j)$ in $\mathcal{K}$. As a result, the kernel matrix for the training data is $K = \Psi^\top \mathcal{J} \Psi$. 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 j-th 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 $\mathbf{z} := [\frac{1}{n_1} \mathbf{1}_{n_1}^\top, -\frac{1}{n_2} \mathbf{1}_{n_2}^\top]^\top$ is an $n \times 1$ vector and $M := (K\mathbf{z})(K\mathbf{z})^\top$, we have

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

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

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

Since N is positive semidefinite and singular, 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 is $b = -\frac{1}{2} \boldsymbol{\alpha}^\top K \mathbf{z}_+$, where $\mathbf{z}_+ := [\frac{1}{n_1} \mathbf{1}_{n_1}^\top,\ \frac{1}{n_2} \mathbf{1}_{n_2}^\top]^\top$. Hence, given that $\mathbf{k}_x = [k(x_1, x), \ldots, k(x_n, x)]^\top$, IKFD is defined as

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

By comparing IKFD to KFD [18], [19], 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 (7) is applicable for indefinite kernels and has a geometric foundation and geometric interpretation in indefinite spaces. Details on IKFD can be found in [13].

IV. 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

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

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. [6] 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 quadratic discriminant functions f_j, $j = 1,\dots,c$

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

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 can be estimated by $P(\omega_j) := n_j/n$.

As discussed in [15], 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 e.g. $T = C$ is the empirical covariance operator. This is our motivation for studying quadratic classifiers based on Mahalanobis distances in the implicit kernel feature space.

Hence, in order to describe Kernel Quadratic Discriminant (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 (10) can be expressed by operations on the kernel only, but it will get another treatment in Sec. IV-D in order to elegantly avoid numerical difficulties.

We will now derive three approaches to kernel quadratic discriminant denoted: KQD-IC for *Invertible Covariance operators,* KQD-RC for *Regularized Covariance operators* and KQD-FK for *Full Kernel matrix*. Each of these methods has a proper extension to indefinite kernels yielding IKQD-IC, IKQD-RC and IKQD-FK, respectively. Different regularization methods are indicated by additional sub-/superscripts. In particular, superscript $^+$ indicates regularization by a suitable *addition*, while superscript $^-$ indicates regularization by a suitable *removal* (or *simplification*) step.

A. 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 (non-singular) empirical class covariance operators C in the kernel induced space. This limits our reasoning to a finite-dimensional $\mathcal{H}$, as the image of an empirical covariance operator C based on n samples has a finite dimension $m < n$. The following considerations require identical class-wise derivations. Therefore, we concentrate on a single class of n elements $\Phi = [\phi(x_1),\dots,\phi(x_n)]$ and drop the super-/subscript j to simplify the notation. Remember that the empirical mean of Φ is $\phi_\mu := \frac{1}{n}\Phi\mathbf{1}_n$, the centered configuration is $\tilde{\Phi} := \Phi - \phi_\mu\mathbf{1}_n^\mathsf{T} = \Phi H$ and the invertible (thanks to our assumption) empirical covariance operator is $C := \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T}$. 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 I. Since $\mathcal{H}$ is m-dimensional, with $m < n$, we may interpret $\tilde{\Phi}$ as an $m \times n$ matrix. Hence, it has a singular value decomposition $\tilde{\Phi} = USV^\mathsf{T}$ with orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $S \in \mathbb{R}^{m\times n}$. By using the orthogonality of U and V, we have: $C = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T} = \frac{1}{n}USS^\mathsf{T}U^\mathsf{T}$ and $\tilde{K} = \tilde{\Phi}^\mathsf{T}\tilde{\Phi} = VS^\mathsf{T}SV^\mathsf{T}$, with an invertible matrix $SS^\mathsf{T} \in \mathbb{R}^{m\times m}$ but singular $S^\mathsf{T}S \in \mathbb{R}^{n\times n}$. So $C^{-1} = nU(SS^\mathsf{T})^{-1}U^\mathsf{T}$ and $\tilde{K}^- = V(S^\mathsf{T}S)^-V^\mathsf{T}$, where the superscript $^-$ denotes the pseudo-inverse of a matrix. Multiplication of these equations with $\tilde{\Phi}$ yields

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

Since $S \in \mathbb{R}^{m\times n}$ is diagonal and has m nonzero singular values, both middle matrices $S(S^\mathsf{T}S)^-$ and $(SS^\mathsf{T})^{-1}S$ 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:

$$C\tilde{\phi}(x) = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T}\Big(\phi(x) - \frac{1}{n}\Phi\mathbf{1}_n\Big) = \frac{1}{n}\tilde{\Phi}H\Big(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n\Big) = \frac{1}{n}\tilde{\Phi}\tilde{\mathbf{k}}_x,$$

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

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

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

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

In a c-class problem, a quadratic discriminant for the j-th class is obtained from (9) by inserting our estimated quantities as

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

$\tilde{K}^{[j]}$ is singular as $\operatorname{rank}(\tilde{K}^{[j]}) < n_j$ due to kernel centering. Hence, we can rely on the pseudo-inverse of $\tilde{K}^{[j]}$ with a given tolerance $\alpha_j > 0$, which means that singular values smaller than

α_j are treated as 0. Alternatively, we can use the inverse of a regularized kernel $\tilde{K}^{[j]}_{\text{reg}} := \tilde{K}^{[j]} + \alpha_j I_{n_j}$, where $\alpha_j > 0$ is a small regularization constant. This leads to alternative discriminant functions of the form

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

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

IKQD-IC, extension to indefinite kernels

We again assume non-singular empirical class covariance operators of data embedded into a finite-dimensional Kreĭn 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 I-A for details. As a result, we obtain the following quadratic discriminants for the IKQD-IC$^-$approach:

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

This is equivalent to (14) except for the definiteness of the kernel matrix. The inverse of regularized $\tilde{K}^{[j]}$ can again be used instead of the pseudo-inverse $(\tilde{K}^{[j]})^-$. Remember that regularizing $\tilde{K}^{[j]}$ by adding a constant α_j to the diagonal, $\tilde{K}^{[j]} + \alpha_j I_{n_j}$, is equivalent to enlarging the original eigenvalues of $\tilde{K}^{[j]}$ by α_j. $\tilde{K}^{[j]}$ is however an indefinite kernel that has both positive and negative eigenvalues. Regularization should therefore be in agreement with this property. Let us consider an eigendecomposition $\tilde{K}^{[j]} = U^{[j]}\Lambda^{[j]}(U^{[j]})^\mathsf{T}$, where $\Lambda^{[j]} = \text{diag}(\boldsymbol{\lambda}_+^{[j]}, \boldsymbol{\lambda}_-^{[j]}, \mathbf{0})$ 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 $J^{[j]} := \text{diag}(\mathbf{1}_{p_j}, -\mathbf{1}_{q_j}, \mathbf{1}_{n_j-p_j-q_j})$, we imply that $\Lambda^{[j]} = |\Lambda^{[j]}|J^{[j]}$. We can then easily verify that $\tilde{K}^{[j]} = U^{[j]}|\Lambda^{[j]}|J^{[j]}(U^{[j]})^\mathsf{T}$. Hence, we will therefore define $\tilde{K}^{[j]}_{\text{reg}} := U^{[j]}\Lambda^{[j]}_{\text{reg}}(U^{[j]})^\mathsf{T}$, where $\Lambda^{[j]}_{\text{reg}} := \Lambda^{[j]} + \alpha_j J^{[j]}$, and $\alpha_j > 0$ being a chosen constant. This leads to the IKQD-IC$^+$discriminants expressed as:

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

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

B. 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 $C^{[j]}_{\text{reg}} := \frac{1}{n_j}\tilde{\Phi}^{[j]}(\tilde{\Phi}^{[j]})^\mathsf{T} + \sigma_j^2 I$ 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 I-B for details. In our c-class problem, a quadratic discriminant for the j-th class is therefore defined as:

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

We refer to this method as KQD-RC$^+$(Kernel Quadratic Discriminant with Regularized Covariance operators). There is no need to use a pseudo-inverse 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. Hence, $(\tilde{K}^{[j]}_{\text{reg}})^{-1}$ 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$^-$:

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

IKQD-RC, extension to indefinite kernels

Similarly to the pd case, we deal with finite samples in a high-/infinite-dimensional Kreĭn 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. positive or negative subspaces. The derivations in Appendix I-C are based on the choice $\tilde{K}^{[j]}_{\text{reg}} := \tilde{K}^{[j]} + n_j\sigma_j^2 U^{[j]}J^{[j]}(U^{[j]})^\mathsf{T}$ where $\tilde{K}^{[j]} = U^{[j]}\Lambda^{[j]}(U^{[j]})^\mathsf{T}$ is eigendecomposition of the centered kernel submatrix for the j-th 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 j-th class as:

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

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 $n_j\sigma_j^2 U^{[j]}J^{[j]}(U^{[j]})^\mathsf{T}$. Hence, $(\tilde{K}^{[j]}_{\text{reg}})^{-1} = \frac{1}{n_j\sigma_j^2}U^{[j]}J^{[j]}(U^{[j]})^\mathsf{T}$. This leads to the IKQD-RC$^-$discriminants

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

Note $U^{[j]}J^{[j]}(U^{[j]})^\mathsf{T}$ is *not* diagonal, in contrast to KQD-RC$^-$.

C. KQD-FK, derived in the complete kernel space

Both KQD approaches considered so far build discriminant functions f_j in class-wise subspaces. The functions f_j rely on the class kernel matrices $K^{[j]}$, which means that the between-class information is unused in the Mahalanobis distances. The third approach we want to propose is to define KQD in a complete kernel space. This can be practically realized in a kernel PCA (KPCA) space as derived in detail in Appendix I-D. Let $\tilde{K} = [\tilde{K}_1, \ldots, \tilde{K}_c]$ be the centered kernel matrix for all training objects, where 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 class-wise centered matrices $\tilde{K}^{[j]} \in \mathbb{R}^{n_j\times n_j}$ from KQD-IC. The kernelized Mahalanobis distance is based on the matrix $\tilde{\tilde{K}}^{[j]} := \tilde{K}_j H^{[j]}\tilde{K}_j^\mathsf{T} \in \mathbb{R}^{n\times n}$. Since rank$(\tilde{\tilde{K}}^{[j]}) < n_j$ by construction, its inverse cannot be derived. In analogy to KQD-IC, we either use a pseudo-inverse of $\tilde{\tilde{K}}^{[j]}$ or regularize it by diagonal addition. This leads to the following discriminant

TABLE II

KQD and IKQD approaches for a c-class problem based on decision functions f_j, $j = 1, \dots, c$. 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. Without loss of generality we use Ψ and * (for the conjugate-transpose) to accommodate for both Hilbert and Kreĭn spaces. pinv(A, α) denotes a pseudo-inverse of the matrix A, treating singular values smaller than $|\alpha|$ as zero. $\tilde{K}^{[j]} = U^{[j]}|\Lambda^{[j]}|J^{[j]}(U^{[j]})^{\mathsf{T}}$ 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 $J^{[j]} := \text{diag}(\mathbf{1}_{p_j}, -\mathbf{1}_{q_j}, \mathbf{1}_{n_j - p_j - q_j})$.

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

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$^-$**
$f_j(x) = -\frac{n_j}{2} (\tilde{\tilde{\mathbf{k}}}_x^{[j]})^{\mathsf{T}} (\tilde{\tilde{K}}_{\text{reg}}^{[j]})^{-1} \tilde{\tilde{\mathbf{k}}}_x^{[j]} + b_j$ $\tilde{\tilde{K}}_{\text{reg}}^{[j]} = \tilde{\tilde{K}}^{[j]} + \alpha_j I_n$	$f_j(x) = -\frac{n_j}{2} (\tilde{\tilde{\mathbf{k}}}_x^{[j]})^{\mathsf{T}} (\tilde{\tilde{K}}^{[j]})^- \tilde{\tilde{\mathbf{k}}}_x^{[j]} + b_j$ $(\tilde{\tilde{K}}^{[j]})^- = \text{pinv}(\tilde{\tilde{K}}^{[j]}, \alpha_j)$

functions, denoted as KQD-FK$^-$(Kernel Quadratic Discriminant in the Full Kernel space),

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

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

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

IKQD-FK, extension to indefinite kernels

We again denote $\tilde{K} = [\tilde{K}_1, \dots, \tilde{K}_c]$ as the centered kernel matrix for all training objects, where 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) [22] allows one to derive the kernel Mahalanobis distance based on the matrix $\tilde{\tilde{K}}^{[j]} := \tilde{K}_j H^{[j]} \tilde{K}_j^{\mathsf{T}} \in \mathbb{R}^{n \times n}$ as worked out in Appendix I-E. Again, rank$(\tilde{\tilde{K}}^{[j]}) < n_j$ by construction, so its inverse cannot be computed. We can either use a pseudo-inverse of $\tilde{\tilde{K}}^{[j]}$ or regularize it appropriately. Note that independently of the definiteness of $\tilde{K}$, $\tilde{\tilde{K}}^{[j]}$ is always positive semidefinite because it is an inner product matrix $\tilde{\tilde{K}}^{[j]} = (\tilde{K}_j H^{[j]})(\tilde{K}_j H^{[j]})^{\mathsf{T}}$. Consequently, both pseudo-inverse and regularization by diagonal addition work identically as in the positive definite case. This means that the IKQD-FK$^-$discriminants are described by formula (22), while the IKQD-FK$^+$discriminants are expressed by formula (23). The difference again only lies in the definiteness of the kernel. The summary of all KQD approaches is presented in Table II.

D. Choice of bias

As stated before, it is possible to derive the bias b_j in the discriminant function f_j of a MAP decision only by 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 non-singular covariance matrix $C^{[j]}$ in a 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. [28]. 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 $i = 1, \dots, l := \text{rank}(\tilde{K}^{[j]})$. Similar procedures can also be derived for regularized covariance matrices.

However, numerical problems arise because a 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

TABLE III

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	$\mathcal{O}(pn^3 + cp^3)$	$\mathcal{O}(np + cp^2)$

for 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 this instability with respect to the estimated intrinsic dimension, this resulting (unconfident) bias b_j experimentally turns out to dominate the Mahalanobis distance contribution frequently. As a result, it spoils the predictability of the resulting classification rule. Therefore, we apply another interpretation of the bias values b_j in the traditional QDA, which leads to a stable and elegant computation scheme for the biases of kernelized classifiers.

In case of class-wise normally distributed data, the traditional QDA (with exact mean and covariance) is the Bayes classifier [6]. 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, 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. Fixing 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 these two classes. For a c-class problem, this can be applied in a class-wise 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

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

E. Computational complexity

Table III 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 denoted by 'I' in the front of the used abbreviations. We assume a c-class problem with n training samples. For simplicity, we assume equal class-priors and set $n_j := n/c$ 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 class-wise 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. 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 Sec. IV-D. The computation of c (pseudo-)inverses matrix 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 here only auxiliary vectors need to be computed in $\mathcal{O}(cn_j^2)$. This gives the training complexities as shown in the left column. 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 exponent $\alpha \geq 3$ would be realistic for general QP solvers).

Observe that the KQD-IC and KQD-RC approaches are clearly beneficial in 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 non-sparse 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.

F. Related methods

Various nonlinear kernel-based techniques, including the kernel Mahalanobis distance, are considered in [27]. As such, the pure kernel Mahalanobis distance (KMD) is used there in a class-wise manner. This is analogous to our KQD-IC$^-$approach relying on the pseudo-inverse of the class kernel submatrix but without the use of bias values. The authors report a good performance of KMD on RBF kernels defined for some standard vectorial data from Machine Learning Repository.

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 [24]. 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 $f(x) = w^\intercal \phi(x)$. 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 the 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

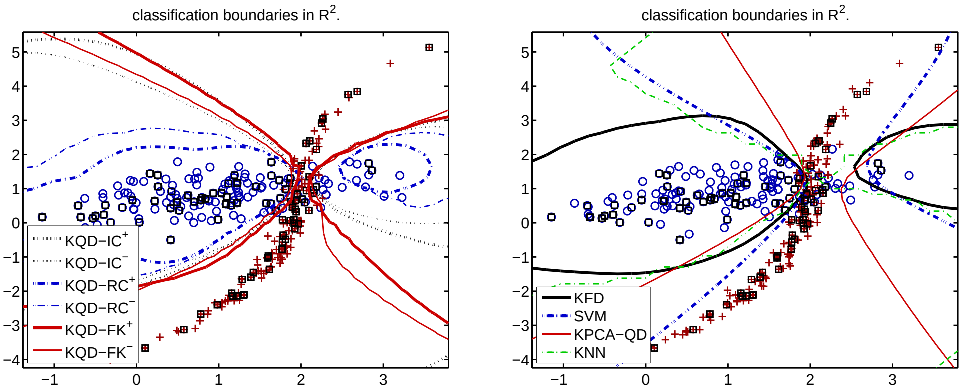

Fig. 1. Classifiers on 2D toy data based on the Gaussian RBF-kernel. The classifier parameters are determined via cross-validation. The left plot shows results for all KQD methods, while the right plot shows results for the reference methods.

TABLE IV

AVERAGE TEST ERRORS [IN %] OF DIFFERENT KERNEL-BASED CLASSIFIERS FOR THE RBF KERNEL ON 2D NONSEPARABLE DATA. THE KERNEL PARAMETER s VARIES. AVERAGING IS PERFORMED OVER TEN DATA DRAWINGS.

Classifier	Kernel parameter s						
	0.5	1	5	10	50	100	500
KQD-IC$^+$	36.3 (8.4)	12.3 (1.9)	6.6 (0.7)	6.8 (1.0)	6.9 (1.3)	7.4 (1.3)	8.6 (2.3)
KQD-IC$^-$	37.7 (8.4)	11.6 (1.4)	6.6 (0.7)	6.7 (1.0)	7.0 (0.8)	6.9 (1.0)	9.5 (3.0)
KQD-RC$^+$	10.7 (1.4)	10.0 (3.1)	14.3 (1.7)	14.6 (4.5)	16.8 (3.3)	12.8 (1.8)	22.8 (3.6)
KQD-RC$^-$	10.9 (2.0)	11.8 (2.2)	17.3 (5.8)	16.1 (4.3)	15.0 (5.9)	17.6 (4.6)	22.7 (3.7)
KQD-FK$^+$	9.5 (1.1)	7.8 (0.7)	6.4 (1.1)	6.4 (0.5)	8.4 (1.6)	9.1 (2.0)	10.1 (2.2)
KQD-FK$^-$	9.6 (1.2)	7.8 (0.8)	6.7 (1.1)	6.9 (1.3)	9.0 (2.6)	9.6 (2.0)	10.0 (2.1)
KFD	9.8 (1.8)	10.3 (2.5)	12.7 (2.2)	13.2 (2.5)	16.2 (3.3)	21.9 (3.0)	23.7 (2.6)
SVM	9.9 (1.7)	9.3 (2.0)	7.8 (1.9)	9.2 (2.6)	13.0 (2.4)	13.1 (2.9)	18.9 (3.4)
KNN	10.3 (1.3)	10.0 (1.3)	10.0 (1.3)	10.0 (1.3)	10.0 (1.3)	10.0 (1.3)	10.0 (1.3)
KPCA-QD	9.4 (1.9)	10.3 (2.8)	11.6 (2.0)	9.7 (2.6)	9.5 (1.4)	9.5 (1.4)	9.5 (1.4)

distribution $p(f_x|X_{\mathrm{tr}},\mathbf{y},x)$ has the mean $\bar{f}_x = \tilde{\mathbf{k}}_x^{\mathsf{T}}(\tilde{K}+\alpha_n^2 I_n)^{-1}\mathbf{y}$ and the variance var$(f_x) = \tilde{k}_{xx} - \tilde{\mathbf{k}}_x^{\mathsf{T}}(\tilde{K}+\alpha_n^2 I_n)^{-1}\tilde{\mathbf{k}}_x$; see [24] for details. Hence, in particular, 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 [15]. In particular, the authors present a statistical support for KFD and they 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 by [15].

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.

V. 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 Sec. IV-E. The reference methods are also applicable for indefinite kernels, cf. [10], [22] and Sec. III-A. Consequently, the reference methods will be denoted ISVM, IKFD, IKNN and IKPCA-QD in case of indefinite kernel matrices. All experiments rely on the MATLAB package PRtools41 [7]. 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 is guaranteed to converge for indefinite kernels.

A. *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 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 eight values on a logarithmic scale. The value $k \in \{1,\dots,8\}$ is optimized for KNN. Classification results are found on independently drawn test sets of $500+500$ examples, as also reported in Table IV.

For illustration purposes, we select the best kernel parameter s

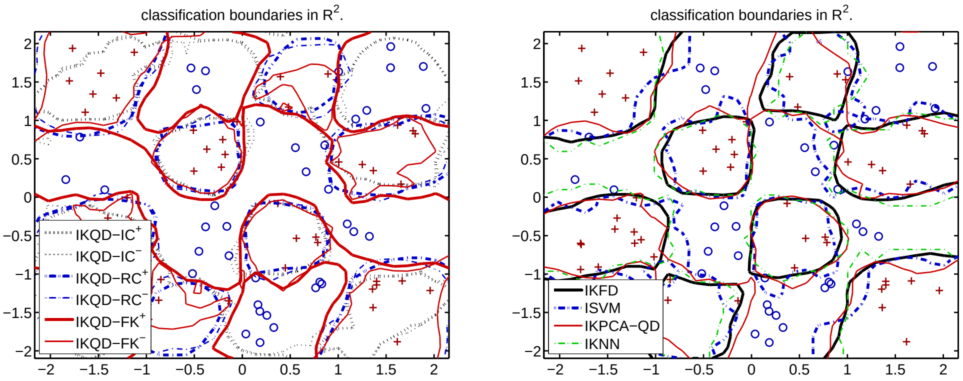

Fig. 2. Classifiers on 2D checkerboard data based on an invariant and indefinite Gaussian RBF-kernel. The classifier parameters are determined via cross-validation. Left plot shows results for all IKQD methods, while right plot shows results for the reference methods.

TABLE V

INDICES OF INDEFINITENESS AND AVERAGE TEST ERRORS FOR DIFFERENT KERNEL-BASED CLASSIFIERS BASED ON THE INVARIANT GAUSSIAN RBF KERNEL FOR CHECKERBOARD DATA. THE KERNEL PARAMETER s VARIES. AVERAGING IS PERFORMED OVER TEN DATA DRAWINGS.

Indefiniteness	Kernel parameter s						
	0.05	0.1	0.5	1	5	10	50
r_{neg}	0.160	0.180	0.211	0.218	0.214	0.207	0.128
(p,q)	(56,44)	(54,46)	(52,48)	(52,48)	(51,49)	(51,49)	(49,51)
$\|\|\phi_\mu^{[1]} - \phi_\mu^{[2]}\|\|_{\mathcal{K}}^2$	0.165	0.180	0.150	0.075	-0.068	-0.027	0.062
Classifier	0.05	0.1	0.5	1	5	10	50
IKQD-IC$^+$	46.1 (3.4)	43.7 (3.6)	16.2 (1.9)	15.0 (2.9)	20.6 (3.9)	21.6 (4.8)	20.7 (3.0)
IKQD-IC$^-$	49.5 (3.9)	39.2 (4.5)	13.7 (2.0)	12.4 (1.8)	17.2 (4.0)	18.7 (2.2)	18.6 (2.8)
IKQD-RC$^+$	15.0 (2.3)	15.0 (3.3)	19.5 (3.0)	30.4 (4.8)	37.5 (7.3)	47.6 (4.2)	35.7 (5.1)
IKQD-RC$^-$	15.3 (2.3)	15.0 (2.3)	13.5 (3.5)	12.9 (3.2)	42.9 (3.7)	50.0 (2.7)	35.4 (5.2)
IKQD-FK$^+$	14.5 (2.5)	15.1 (2.2)	12.5 (1.8)	12.7 (1.8)	13.9 (2.3)	13.5 (2.3)	19.5 (2.0)
IKQD-FK$^-$	21.6 (4.1)	19.3 (4.0)	15.2 (2.5)	17.1 (4.0)	14.9 (2.9)	13.3 (2.9)	20.0 (1.7)
IKFD	14.0 (1.6)	11.7 (2.6)	14.3 (1.9)	13.6 (3.4)	14.6 (3.5)	12.8 (1.3)	21.7 (3.3)
ISVM	20.6 (3.1)	22.8 (3.3)	28.7 (4.6)	45.2 (6.0)	46.1 (6.4)	49.4 (5.8)	38.5 (5.8)
IKNN	15.0 (2.4)	15.4 (2.3)	15.4 (2.3)	15.4 (2.3)	15.4 (2.3)	15.4 (2.3)	15.4 (2.3)
IKPCA-QD	13.7 (1.4)	12.4 (2.6)	13.0 (2.3)	12.0 (3.2)	12.9 (2.4)	12.8 (2.4)	18.6 (3.8)

(by cross-validation) for each classifier. Example KQD-classifiers are depicted in Fig. 1, left plot. The right plot illustrates reference classifiers: KFD, SVM, KPCA-QD 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 assess the statistical significance and independence of kernel parameter choice, we repeat the above data-drawing, cross-validation and test-error determination ten times. This procedure is repeated for different parameters s. We do not select s during the cross-validation as this implicitly changes the data representation in the feature space. Instead, we consider classification performance on the given kernel-feature space representation for different s. The results are presented in Table IV.

We can observe that the KQD-RC$^+$and KQD-RC$^-$approaches are mostly inferior to the reference classifiers if $s > 1$ is fixed. These methods seem to favor smaller values of s, whereas the KQD-IC classifiers are better for larger s. However, KQD-IC and KQD-FK perform mostly better than the reference classifiers for all fixed $s > 1$. Note that KNN, despite of being a kernel classifier, is independent of the choice of s for the RBF-kernel, which can easily be verified. For low s, however, we reach a range where numerical inaccuracies destroy this s-independence.

To analyze classification performance, we address the overall mean and standard deviation of the test errors determined by (s, C, α)-cross-validated classifiers in ten runs. These are 6.8±1.0 (KQD-IC$^+$), 6.7±1.0 (KQD-IC$^-$), 10.3±1.9 (KQD-RC$^+$), 11.7±2.1 (KQD-RC$^-$), 6.9±1.4 (KQD-FK$^+$), 7.1±1.2 (KQD-FK$^-$) for the KQD approaches and 10.5±2.1 (KFD), 8.9±1.8 (SVM), 9.3±2.3 (KPCA-QD), and 10.1±1.4 (KNN) for the reference classifiers. The KQD-IC and KQD-FK nonlinear kernel classifiers seem to perform much better than the linear SVM and KFD. But they are also superior to other nonlinear kernel classifiers KPCA-QD and KNN.

B. 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. The results of experiments are reported in Table V. 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 [21], [11]. We observe that the checkerboard distribution is invariant wrt. 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 [12]. 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 Sec. V-A, i.e. for a fixed kernel value s, a training set of $50+50$ elements is drawn. Parameters of all classifiers are found via 10-fold cross-validation. Test error rates are determined on an independent test set of $500+500$ samples. This is averaged over ten data drawings and repeated for different kernel parameter s. The ranges of s and α are slightly shifted as compared to the previous section.

Example classifiers are illustrated in Fig. 2, where both s and the regularization parameters are determined by 10-fold cross-validation. One can clearly observe the perfect point symmetry of all classifiers thanks to the use of invariant kernel, even though the training set is asymmetric. To maintain the clarity of presentation, the test examples are not plotted.

We assess some measures of the indefiniteness of the resulting kernel matrices. First, we determine the signature (p,q) of the kernel, 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 Kreĭn space $\mathcal{K}$. In addition, Table V provides an index of indefiniteness, $r_{\text{neg}} := \left(\sum_{\lambda_i<0} |\lambda_i|\right) / \left(\sum_i |\lambda_i|\right)$, the ratio of negative variance to overall variance measured by the sums of eigenvalues λ_i of K, and the squared distance of the class means $||\phi_\mu^{[1]} - \phi_\mu^{[2]}||^2_{\mathcal{K}}$. Finally, the corresponding average test errors are also reported there.

Concerning indefiniteness, we note that the fraction of negative energy is the highest in the middle range of s and is decreasing towards both lower and higher values. This is expected, because kernel matrices converge to either I_n for $s \to 0$ or to the matrix $\mathbf{1}_n\mathbf{1}_n^\top$ for $s \to \infty$, which are both positive semidefinite. Note that the square distance between class means in the embedded Kreĭn space may by negative for some s. This gives rise to difficult separation with indefinite SVM [10]. Indeed, ISVM performs badly in these cases. We again observe that both IKQD-RC methods seem to favor smaller s values, while the IKQD-IC approaches give better results for larger s. In order to compare classification performance, we address the overall mean and standard deviation of the test errors determined by (s, C, α)-cross-validated classifiers in ten runs. These are 16.7±3.3 (IKQD-IC$^+$), 13.5±2.4 (IKQD-IC$^-$), 14.8±1.7 (IKQD-RC$^+$), 14.2±2.3 (IKQD-RC$^-$) 12.9±1.9 (IKQD-FK$^+$), 14.3±3.6 (IKQD-FK$^-$) for the IKQD approaches and 13.2±2.1 (IKFD), 19.6±4.5 (ISVM), 14.6±4.1 (IKPCA-QD), and 16.7±2.8 (IKNN). In conclusion, we see that all IKQD-approaches outperform ISVM and IKNN and all except IKQD-IC$^+$and IKQD-RC$^+$are slightly superior to IKPCA-QD. Finally, IKQD-FK$^+$is on average better than IKFD.

C. Real-world kernel data

We now consider both two-class and multi-class 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.

TABLE VI

CHARACTERISTICS OF REAL-WORLD KERNEL DATA. β IS THE FRACTION OF DATA USED FOR TRAINING IN THE HOLD-OUT EXPERIMENTS. THE INDICES OF INDEFINITENESS OF THE KERNEL, $r_{\text{NEG}} \in [0,1]$ AND (p,q), ARE AVERAGED OVER 25 RUNS.

	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 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. This is only important for practical reasons in order to use identical ranges of parameters in cross-validation for different datasets.

The centered training kernel matrix $\tilde{K}$ obtained from a dissimilarity function is pd *only if* the dissimilarity matrix $D := (d(x_i,x_j))_{i,j=1}^n$ is isometrically embedabble into a Euclidean space [9], [21]. 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 IKQD and reference classifiers.

The data sets are described in Appendix II, while kernel matrices are briefly characterized in Table VI. 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 class-wise centering).

We run hold-out experiments in which the complete data is split into the training and test kernel matrices such that the specified β-fraction is used for training; see Table VI. In each run, parameters of all classifiers are determined by 10-fold cross-validation. Here, the class-wise regularization parameters are 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, $C \in [10^{-1}, 10^8]$ for ISVM. The total number of investigated values

TABLE VII

AVERAGE CLASSIFICATION ERRORS [IN %] FOR POSITIVE-DEFINITE AND INDEFINITE KERNEL DATA. NUMBERS IN PARENTHESIS DENOTE STANDARD DEVIATIONS. BEST IKQD AND REFERENCE CLASSIFIER ARE HIGHLIGHTED IN EACH COLUMN.

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)

is $11-13$. For IKNN, the value $k \in \{1,2,\dots,45\}$ 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 complete procedure is repeated 25 times for all classifiers and the results are averaged.

Table VII shows average classification errors and standard-deviations for the IKQD classifiers and reference methods. Problems with nearly pd kernels are: *Nist38-EU* (pd), *Hart* (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. There are two reasons. First, except for the search of biases, the discriminants do not make use of the between-class information. So, they can only work well for 'clean' separation between the classes (as e.g. in the *Nist38-EU* case), which means that the between-class kernel values are much smaller than the within-class kernel values. Secondly, there are numerical difficulties caused by the use of the second power of (pseudo)-inverse of the class-wise kernel submatrix; see (17) and (16). If there is insufficient discriminative information in the class-related kernel, it will be enhanced in this process.
- One of the best methods are usually either IKQD-RC$^+$or IKQD-FK$^+$.

Among the reference methods we see that:

- ISVM performs badly for multi-class 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

- 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.

VI. 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 use. Ignoring the computation of the bias b_j, the approaches KQD-IC and KQD-RC do not use between-class information. In contrast, the KQD-FK methods rely on the full kernel information for the computation of Mahalanobis distance. In addition to the test-versus-train matrix, the KQD-RC approaches require the diagonal kernel values $k(x,x)$. Concerning computation complexities, the KQD-IC and KQD-RC approaches have the conceptual advantage of 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 multi-class 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 Kreĭn spaces. We also propose extensions of all discussed KQD discriminants to indefinite kernels. All the methods have a sound mathematical motivation, hence extend the class of kernel methods that can work with indefinite kernels.

Experimentally, the IKQD-RC$^+$and IKQD-FK$^+$seem to be favorable among the IKQD approaches. The latter seems the most beneficial, but is computationally more expensive due to full kernel matrix processing. The 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, e.g. *Nist38-MH, Poly-H, Chicken-15* and *Chromo-DIF*, or 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 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-wise or full kernel-induced subspaces. More research is however needed to clearly identify conditions under which the nonlinear KQD/IKQD methods will outperform linear KFD/IKFD and SVM/ISVM.

Acknowledgements

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," in *Advances in Learning Theory: Methods, Models and Applications*, J. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, Eds. IOS Press, 2003, vol. 190, pp. 89–110.

[3] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for Support Vector Machines," 2001, available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[4] M. Dritschel and J. Rovnyak, "Operators on indefinite inner product spaces," *Lectures on Operator Theory and its Applications, Fields Institute Monographs*, pp. 141–232, 1996.

[5] M. Dubuisson and A. Jain, "Modified Hausdorff distance for object matching," in *International Conference on Pattern Recognition*, vol. 1, 1994, pp. 566–568.

[6] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, Inc., 2001.

[7] R. Duin, P. Juszczak, D. de Ridder, P. Paclík, E. Pękalska, and D. Tax, "PR-Tools," 2004, http://prtools.org.

[8] L. Goldfarb, "A new approach to pattern recognition," in *Progress in Pattern Recognition*, L. Kanal and A. Rosenfeld, Eds. Elsevier Science Publishers, 1985, vol. 2, pp. 241–402.

[9] J. Gower, "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification*, vol. 3, pp. 5–48, 1986.

[10] B. Haasdonk, "Feature space interpretation of SVMs with indefinite kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 482–492, 2005.

[11] ——, "Transformation knowledge in pattern analysis with kernel methods — distance and integration kernels," Ph.D. dissertation, Universität Freiburg, Institut für Informatik, 2005.

[12] B. Haasdonk and H. Burkhardt, "Invariant kernel functions for pattern analysis and machine learning," *Machine Learning*, vol. 68, no. 1, pp. 35–61, 2007.

[13] B. Haasdonk and E. Pękalska, "Indefinite kernel Fisher discriminant," in *International Conference on Pattern Recognition*, 2008.

[14] S. Hochreiter and K. Obermayer, "Support vector machines for dyadic data," *Neural Computation*, vol. 18, no. 6, pp. 1472–1510, 2006.

[15] S.-Y. Huang, C.-R. Hwang, and M.-H. Lin, "Kernel Fisher's discriminant analysis in Gaussian reproducing kernel Hilbert space," Academia Sinica, Taipei, Taywan, Tech. Rep., 2005.

[16] D. Jacobs, D. Weinshall, and Y. Gdalyahu, "Classification with Non-Metric Distances: Image Retrieval and Class Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 583–600, 2000.

[17] J. Laub and K.-R. Müller, "Feature discovery in non-metric pairwise data," *Journal of Machine Learning Research*, pp. 801–818, 2004.

[18] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing*, 1999, pp. 41–48.

[19] S. Mika, A. Smola, and B. Schölkopf, "An improved training algorithm for kernel Fisher discriminants," in *AISTATS*, 2001, pp. 98–104.

[20] C. Ong, X. Mary, S. Canu, and S. A.J., "Learning with non-positive kernels," in *International Conference on Machine Learning*, Brisbane, Australia, 2004, pp. 639–646.

[21] E. Pękalska and R. Duin, *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, 2005.

[22] ——, “Indefinite kernel PCA,” *work in progress*, 2008.

[23] E. Pękalska, A. Harol, R. Duin, B. Spillmann, and H. Bunke, “Non-Euclidean or non-metric measures can be informative,” in *Joint IAPR Workshops on SSPR and SPR*, 2006, pp. 871–880.

[24] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[25] V. Roth, J. Laub, J. Buhmann, and K.-R. Müller, “Going metric: Denoising pairwise data,” in *Advances in Neural Information Processing Systems*, 2003, pp. 841–856.

[26] J. Rovnyak, “Methods of Krein space operator theory,” *Operator Theory: Advances and Applications*, vol. 134, pp. 31–66, 2002.

[27] A. Ruiz and P. Lopez-de Teruel, “Nonlinear kernel-based statistical pattern analysis,” *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 16–32, 2001.

[28] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge: MIT Press, 2002.

[29] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol. 10, no. 5, 1998.

[30] B. Schölkopf, K. Tsuda, and J. Vert, *Kernel Methods in Computational Biology*. Cambridge, MA: MIT Press, 2004.

[31] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. UK: Cambridge University Press, 2004.

[32] P. Simard, Y. A. Le Cun, J. S. Denker, and B. Victorri, “Transformation invariance in pattern recognition – tangent distance and tangent propagation,” *International Journal of Imaging System and Technology*, vol. 11, no. 3, pp. 181–194, 2001.

[33] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[34] G. Wahba, “Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV,” in *Advances in Kernel Methods, Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999, pp. 69–88.

Appendix I
Derivations

Here we provide derivations details of the IKQD methods.

A. IKQD-IC

Remember that the empirical mean is $\psi_\mu := \frac{1}{n}\Psi\mathbf{1}_n$, the centered configuration is $\tilde{\Psi} = \Psi H$, the empirical covariance operator is $C := \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^\mathsf{T}\mathcal{J}$ and the centered kernel matrix is $\tilde{K} = \tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\Psi}$. As we assume invertibility of C, we implicitly restrict to finite-dimensional spaces $\mathcal{K}$, as the image of C is at most $(n-1)$-dimensional. Hence $\tilde{\Psi}$ is interpreted as a $m \times n$ matrix for $m < n$. Recall that $\tilde{K}$ is singular due to centering. We are now interested in a kernel formulation of the terms $D^2_M(\psi(x); \{\psi_\mu, C\})$ for new observations $x \in \mathcal{X}$. Analogously to the pd case, we have

$$C\tilde{\Psi} = \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\Psi} = \frac{1}{n}\tilde{\Psi}\tilde{K}.$$

Assuming the singular value decomposition $\tilde{\Psi} = USV^T$ with orthogonal $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, diagonal $S = [\bar{S}, \mathbf{0}] \in \mathbb{R}^{m\times n}$ and invertible diagonal $\bar{S} \in \mathbb{R}^{m\times m}$, we see that $SS^T = \bar{S}^2$ is invertible. Since $C^{-1} = nU(SS^\mathsf{T})^{-1}U^\mathsf{T}$ and $\tilde{K}^- = V(S^\mathsf{T}S)^-V^\mathsf{T}$, we arrive at slightly extended terms as compared to (11)

$$\begin{aligned} \frac{1}{n}C^{-1}\tilde{\Psi} &= \mathcal{J}U(SS^\mathsf{T})^{-1}SV^\mathsf{T} = \mathcal{J}U\left[\bar{S}^{-1}, \mathbf{0}\right]V^\mathsf{T} \\ \tilde{\Psi}\tilde{K}^- &= US(S^\mathsf{T}U^\mathsf{T}\mathcal{J}US)^-V^\mathsf{T}. \end{aligned} \tag{24}$$

We aim at the identity of both right hand side terms and indeed by abbreviating $W := U^\mathsf{T}\mathcal{J}U \in \mathbb{R}^{m\times m}$ and, observing that $W = W^{-1}$, we obtain for the middle matrix in the last term

$$S(S^\mathsf{T}WS)^- = [\bar{S}, \mathbf{0}]\begin{bmatrix} \bar{S}^{-1}W^{-1}\bar{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = W^{-1}\left[\bar{S}^{-1}, \mathbf{0}\right].$$

Therefore, the second equation in (24) simplifies to

$$US(S^\mathsf{T}WS)^-V^\mathsf{T} = U(U^\mathsf{T}\mathcal{J}U)\left[\bar{S}^{-1}, \mathbf{0}\right]V^\mathsf{T} = \mathcal{J}U\left[\bar{S}^{-1}, \mathbf{0}\right]V^\mathsf{T}$$

and is equivalent to the first line in (24). Hence, we get the analogue of (12) as:

$$\frac{1}{n}C^{-1}\tilde{\Psi} = \tilde{\Psi}\tilde{K}^-. \tag{25}$$

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

$$C\,\tilde{\psi}(x) = \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^\mathsf{T}\mathcal{J}\left(\psi(x) - \frac{1}{n}\Psi\mathbf{1}_n\right) = \frac{1}{n}\tilde{\Psi}\tilde{\mathbf{k}}_x, \tag{26}$$

where $\mathbf{k}_x = \Psi^\mathsf{T}\mathcal{J}\psi(x)$ and $\tilde{\mathbf{k}}_x = \tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\psi}(x)$. Hence, $\tilde{\psi}(x) = \frac{1}{n}C^{-1}\tilde{\Psi}\tilde{\mathbf{k}}_x = \tilde{\Psi}\tilde{K}^-\tilde{\mathbf{k}}_x$ thanks to (25). Then we can express the empirical Mahalanobis distance as

$$\begin{aligned} D^2_M(\psi(x); \{\psi_\mu, C\}) &= \tilde{\psi}(x)^\mathsf{T}\mathcal{J}C^{-1}\tilde{\psi}(x) \\ &= \tilde{\psi}(x)^\mathsf{T}C^{-1}\tilde{\Psi}\tilde{K}^-\tilde{\mathbf{k}}_x = n\tilde{\mathbf{k}}_x^\mathsf{T}((\tilde{K})^-)^2\tilde{\mathbf{k}}_x. \end{aligned}$$

This yields the decision function (16) of IKQD-IC$^+$.

B. KQD-RC

As an ansatz we directly regularize the covariance operator:

$$C_{\text{reg}} := C + \sigma^2 I_n = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^\mathsf{T} + \sigma^2 I_n, \tag{27}$$

where $\sigma^2 > 0$ is a parameter to be chosen. By multiplying by $\tilde{\Phi}$ from both sides, using $\tilde{K} = \tilde{\Phi}^\mathsf{T}\tilde{\Phi}$ and defining $\tilde{K}_{\text{reg}} := \tilde{K} + \alpha I_n$ for $\alpha := n\sigma^2$, we get

$$C_{\text{reg}}\tilde{\Phi} = \frac{1}{n}\tilde{\Phi}(\tilde{K} + n\sigma^2 I_n) = \frac{1}{n}\tilde{\Phi}\tilde{K}_{\text{reg}}.$$

As a result, both C_{reg} and $\tilde{K}_{\text{reg}}$ are strictly positive definite, in particular non-singular, as $n\sigma^2 > 0$. Hence, the inverses are well-defined and we can equivalently write

$$\tilde{\Phi}\tilde{K}_{\text{reg}}^{-1} = \frac{1}{n}C_{\text{reg}}^{-1}\tilde{\Phi} \tag{28}$$

Note that $nC\tilde{\phi}(x) = \tilde{\Phi}\tilde{\Phi}^\mathsf{T}(\phi(x) - \frac{1}{n}\Phi\mathbf{1}_n) = \tilde{\Phi}H(\mathbf{k}_x - \frac{1}{n}K\mathbf{1}_n) = \tilde{\Phi}\tilde{\mathbf{k}}_x$, where $\mathbf{k}_x$ and $\tilde{\mathbf{k}}_x$ are defined in (2). Hence, C_{reg} acts on an arbitrary centered vector $\tilde{\phi}(x)$ as: $C_{\text{reg}}\,\tilde{\phi}(x) = \frac{1}{n}\tilde{\Phi}\tilde{\mathbf{k}}_x + \sigma^2\tilde{\phi}(x)$. Since C_{reg} is invertible, we obtain

$$\tilde{\phi}(x) = \frac{1}{n}C_{\text{reg}}^{-1}\tilde{\Phi}\tilde{\mathbf{k}}_x + \sigma^2 C_{\text{reg}}^{-1}\tilde{\phi}(x). \tag{29}$$

By multiplying (29) on both sides by $\tilde{\phi}(x)^\mathsf{T}$ (from the left) and thanks to (28), we can write

$$\tilde{\phi}(x)^\mathsf{T}\tilde{\phi}(x) = \tilde{\phi}(x)^\mathsf{T}\tilde{\Phi}\tilde{K}_{\text{reg}}^{-1}\tilde{\mathbf{k}}_x + \sigma^2\tilde{\phi}(x)^\mathsf{T}C_{\text{reg}}^{-1}\tilde{\phi}(x).$$

By abbreviating $\tilde{k}_{xx} = \tilde{\phi}(x)^\mathsf{T}\tilde{\phi}(x)$ and solving for the last term, we get the desired square Mahalanobis distance $\tilde{\phi}(x)^\mathsf{T}C_{\text{reg}}^{-1}\tilde{\phi}(x) = \frac{1}{\sigma^2}(\tilde{k}_{xx} - \tilde{\mathbf{k}}_x^\mathsf{T}K_{\text{reg}}^{-1}\tilde{\mathbf{k}}_x)$. This leads to the class-wise decision function (18).

C. IKQD-RC

Regularization of an empirical covariance operator in Kreĭn spaces should respect definiteness of the space. Therefore, we set

$$C_{\text{reg}} := C + \sigma^2\mathcal{J} = \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^* + \sigma^2\mathcal{J}, \tag{30}$$

where $\sigma^2 > 0$ is a parameter to be chosen. After multiplying by $\tilde{\Psi}$ from both sides and thanks to $\tilde{K} = \tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\Psi}$, we get

$$C_{\text{reg}}\tilde{\Psi} = \frac{1}{n}\tilde{\Psi}\tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\Psi} + \sigma^2\mathcal{J}\tilde{\Psi} = \frac{1}{n}\tilde{\Psi}\tilde{K} + \sigma^2\mathcal{J}\tilde{\Psi}. \tag{31}$$

We want to express the right-hand side of (31) as $\frac{1}{n}\tilde{\Psi}\tilde{K}_{\text{reg}}$ for some specifically regularized kernel. Therefore we must rewrite the last term $\mathcal{J}\tilde{\Psi}$. Assume an eigendecomposition of $\tilde{\Psi}^*\tilde{\Psi} = \tilde{K}$ as $\tilde{\Psi}^*\tilde{\Psi} = U\Lambda U^\mathsf{T}$ such that $\Lambda = \text{diag}(\boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-, \mathbf{0})$ is a diagonal matrix consisting of p positive, q negative and $(n-p-q)$ zero eigenvalues, while the corresponding eigenvectors are stored in the columns of $U = [U_+, U_-, U_0] \in \mathbb{R}^{n\times n}$. To simplify the following reasoning we assume a specific feature space embedding Ψ of the data, which will "cancel out" later on due to the kernelization. So, the explicit centered feature space embedding is $\tilde{\Psi} := \Psi := |\text{diag}(\boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-)|^{\frac{1}{2}}[U_+, U_-]^\mathsf{T}$ into a finite-dimensional Kreĭn space $\mathcal{K} := \mathbb{R}^{(p,q)}$. We then easily verify that indeed $\tilde{\Psi}^*\tilde{\Psi} = [U_+, U_-]\text{diag}(\boldsymbol{\lambda}_+, \boldsymbol{\lambda}_-)[U_+, U_-]^\mathsf{T} = U\Lambda U^\mathsf{T}$ and furthermore $\tilde{\Psi}^\mathsf{T}\tilde{\Psi} = U|\Lambda|U^\mathsf{T}$. We introduce $J := \text{diag}(\mathbf{1}_p, -\mathbf{1}_q, \mathbf{1}_{n-p-q})$ which implies $\Lambda = |\Lambda|J$. Thanks to orthogonality of U, we get

$$\tilde{\Psi}^\mathsf{T}\mathcal{J}\tilde{\Psi} = U|\Lambda|U^\mathsf{T}UJU^\mathsf{T} = \tilde{\Psi}^\mathsf{T}\tilde{\Psi}UJU^\mathsf{T}. \tag{32}$$

As $\tilde{\Psi}$ spans $\mathcal{K}$, we conclude that $\mathcal{J}\tilde{\Psi} = \tilde{\Psi}UJU^\top$ because they are not discriminating within the span of $\tilde{\Psi}$, i.e. all inner products with $\tilde{\Psi}$ are identical. As a result, (31) becomes

$$C_{\text{reg}}\tilde{\Psi} = \frac{1}{n}\tilde{\Psi}(\tilde{K} + n\sigma^2 UJU^\top) = \frac{1}{n}\tilde{\Psi}\tilde{K}_{\text{reg}},$$

where $\tilde{K}_{\text{reg}} = \tilde{K} + n\sigma^2 UJU^\top = U(\Lambda + n\sigma^2 J)U^\top$. Consequently, $\tilde{K}_{\text{reg}}$ is invertible because the magnitudes of all eigenvalues are enlarged by $n\sigma^2$ and therefore nonzero. This allows us to write

$$\tilde{\Psi}\tilde{K}_{\text{reg}}^{-1} = \frac{1}{n}C_{\text{reg}}^{-1}\tilde{\Psi}.$$

Similarly as in (26), C_{reg} acts on an arbitrary centered vector $\tilde{\psi}(x)$ such that $C_{\text{reg}}\,\tilde{\psi}(x) = \frac{1}{n}\tilde{\Psi}\tilde{\mathbf{k}}_x + \sigma^2\tilde{\psi}(x)$, where $\tilde{\mathbf{k}}_x = \tilde{\Psi}^\top\mathcal{J}\tilde{\psi}(x)$. Since C_{reg} is invertible, we get $\tilde{\psi}(x) = \frac{1}{n}C_{\text{reg}}^{-1}\tilde{\Psi}\tilde{\mathbf{k}}_x + \sigma^2(C_{\text{reg}})^{-1}\tilde{\psi}(x)$. By multiplying with $\tilde{\psi}(x)^\top\mathcal{J}$ from both sides on the left and thanks to (33), we observe that $\tilde{k}_{xx} = \tilde{\psi}(x)^*\tilde{\psi}(x) = \tilde{\psi}(x)^\top\mathcal{J}\tilde{\Psi}\tilde{K}_{\text{reg}}^{-1}\tilde{\mathbf{k}}_x + \sigma^2\tilde{\psi}(x)^\top\mathcal{J}C_{\text{reg}}^{-1}\tilde{\psi}(x)$. By resolving for the last term, the desired square Mahalanobis distance can be expressed as $\tilde{\psi}(x)^\top\mathcal{J}C_{\text{reg}}^{-1}\tilde{\psi}(x) = \frac{1}{\sigma^2}(\tilde{k}_{xx} - \tilde{\mathbf{k}}_x^\top K_{\text{reg}}^{-1}\tilde{\mathbf{k}}_x)$, which yields the class-wise decision function (20).

D. KQD-FK

Let $\tilde{K} = [\tilde{K}_1, \ldots, \tilde{K}_c]$ be a centered kernel matrix for the training data, where 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 here to avoid confusion with the class-wise centered matrices $\tilde{K}^{[j]} \in \mathbb{R}^{n_j\times n_j}$ from KQD-IC. Consider now a vector space extracted via the kernel PCA (KPCA) [29]. We determine an eigendecomposition $\tilde{K} = Q\Lambda Q^\top$, where $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix with r nonzero eigenvalues, and $Q \in \mathbb{R}^{n\times r}$ is the matrix of column-wise orthonormal eigenvectors. The KPCA mapping is defined by $\psi(x) := Q_s^\top\tilde{\mathbf{k}}_x$, where $Q_s := Q\Lambda^{-\frac{1}{2}}$. The explicit KPCA space representation of the j-th class is then denoted as $\Psi^{[j]} := Q_s^\top\tilde{K}_j$, which leads to the empirical mean vector $\psi_\mu^{[j]} := \frac{1}{n_j}\Psi^{[j]}\mathbf{1}_{n_j} = \frac{1}{n_j}Q_s^\top\tilde{K}_j\mathbf{1}_{n_j}$. Consequently, $\tilde{\psi}^{[j]}(x) := \psi(x) - \psi_\mu^{[j]} = Q_s^\top(\tilde{\mathbf{k}}_x - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}) = Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}$, where $\tilde{\tilde{\mathbf{k}}}_x^{[j]} := \tilde{\mathbf{k}}_x - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}$. For the complete j-th class we get

$$\tilde{\Psi}^{[j]} := \Psi^{[j]} - \psi_\mu^{[j]}\mathbf{1}_{n_j}^\top = Q_s^\top\left(\tilde{K}_j - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}\mathbf{1}_{n_j}^\top\right) = Q_s^\top\tilde{K}_jH^{[j]}.$$

This yields the class covariance matrix in the KPCA space as:

$$C^{[j]} = \frac{1}{n_j}\tilde{\Psi}^{[j]}(\tilde{\Psi}^{[j]})^\top = \frac{1}{n_j}Q_s^\top\tilde{K}_jH^{[j]}\tilde{K}_j^\top Q_s = \frac{1}{n_j}Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s,$$

where $\tilde{\tilde{K}}^{[j]} := \tilde{K}_jH^{[j]}\tilde{K}_j^\top \in \mathbb{R}^{n\times n}$ thanks to $H^{[j]} = H^{[j]}H^{[j]}$. Note that $\tilde{K}_j$ is a submatrix of the centered kernel matrix $\tilde{K}$, while $\tilde{\tilde{K}}^{[j]}$ involves additional centering with respect to the j-th class. Given a regular matrix $C^{[j]}$, the empirical square Mahalanobis distance in the KPCA space is

$$\begin{aligned} D_M^2(\psi(x);\{\psi_\mu^{[j]}, C^{[j]}\}) &= (\psi(x) - \psi_\mu^{[j]})^\top(C^{[j]})^{-1}(\psi(x) - \psi_\mu^{[j]}) \\ &= n_j\Big(Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}\Big)^\top\Big(Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s\Big)^{-1}Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]} \\ &= n_j(\tilde{\tilde{\mathbf{k}}}_x^{[j]})^\top Q_s(Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s)^{-1}Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}. \end{aligned}$$

We now motivate our method by a simplification step in order to remove the KPCA dependency. Formally, we propose to remove the matrices Q_s from the above Mahalanobis distance. This would naturally occur if $\tilde{K}$ was regular (which is not true due to kernel centering). With regular $\tilde{K}$, Q_s and Λ would be regular $n \times n$ matrices satisfying $Q_s^\top Q_s = \Lambda^{-1}$. Hence, $Q_s^{-1} = \Lambda Q_s^\top = \Lambda^{\frac{1}{2}}Q$ and $(Q_s^\top)^{-1} = Q_s\Lambda$. In such a case, straightforward computation shows that Q_s cancels out in the Mahalanobis distance D_M^2, which is greatly simplified to $D_M^2(\psi(x);\{\psi_\mu^{[j]}, C^{[j]}\}) = n_j\tilde{\tilde{\mathbf{k}}}_x^{[j]\top}(\tilde{\tilde{K}}^{[j]})^{-1}\tilde{\tilde{\mathbf{k}}}_x^{[j]}$. This leads to decision functions (22) and (23).

E. IKQD-FK

We consider now a vector space extracted from indefinite kernel PCA (IKPCA) [22]. We determine an eigendecomposition $\tilde{K} = Q\Lambda Q^\top$, where $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix with r nonzero eigenvalues, including p positive and q negative ones. $Q \in \mathbb{R}^{n\times r}$ is the matrix of orthonormal eigenvectors. The IKPCA mapping is defined by $\psi(x) := Q_s^\top\tilde{\mathbf{k}}_x$, where $Q_s := Q|\Lambda|^{-\frac{1}{2}}$. The explicit IKPCA space representation of the j-th class is denoted as $\Psi^{[j]} := Q_s^\top\tilde{K}_j$, which leads to the empirical mean vector $\psi_\mu^{[j]} := \frac{1}{n_j}\Psi^{[j]}\mathbf{1}_{n_j} = \frac{1}{n_j}Q_s^\top\tilde{K}_j\mathbf{1}_{n_j}$. Consequently, $\tilde{\psi}^{[j]}(x) := \psi(x) - \psi_\mu^{[j]} = Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}$, where $\tilde{\tilde{\mathbf{k}}}_x^{[j]} := \tilde{\mathbf{k}}_x - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}$. Equivalently, $\tilde{\Psi}^{[j]} := \Psi^{[j]} - \psi_\mu^{[j]}\mathbf{1}_{n_j}^\top = Q_s^\top\left(\tilde{K}_j - \frac{1}{n_j}\tilde{K}_j\mathbf{1}_{n_j}\mathbf{1}_{n_j}^\top\right) = Q_s^\top\tilde{K}_jH^{[j]}$. This leads to the empirical class covariance matrix in the IKPCA space as:

$$C^{[j]} = \frac{1}{n_j}\tilde{\Psi}^{[j]}(\tilde{\Psi}^{[j]})^* = \frac{1}{n_j}Q_s^\top\tilde{K}_jH^{[j]}\tilde{K}_j^\top Q_s\mathcal{J} = \frac{1}{n_j}Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s\mathcal{J},$$

where $\tilde{\tilde{K}}^{[j]} := \tilde{K}_jH^{[j]}\tilde{K}_j^\top \in \mathbb{R}^{n\times n}$ and $H^{[j]} = H^{[j]}H^{[j]}$. Given a regular matrix $C^{[j]}$, the empirical D_M^2 in the IKPCA space is

$$\begin{aligned} D_M^2(\psi(x);\{\psi_\mu^{[j]}, C^{[j]}\}) &= (\psi(x) - \psi_\mu^{[j]})^*(C^{[j]})^{-1}(\psi(x) - \psi_\mu^{[j]}) \\ &= n_j\Big(Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}\Big)^\top\mathcal{J}\Big(Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s\mathcal{J}\Big)^{-1}Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]} \\ &= n_j(\tilde{\tilde{\mathbf{k}}}_x^{[j]})^\top Q_s(Q_s^\top\tilde{\tilde{K}}^{[j]}Q_s)^{-1}Q_s^\top\tilde{\tilde{\mathbf{k}}}_x^{[j]}, \end{aligned}$$

thanks to the identities $\mathcal{J}^{-1} = \mathcal{J}$ and $I_k = \mathcal{J}\mathcal{J}$. We can now simplify this method and remove the dependency on the IKPCA space by removing Q_s from the Mahalanobis distance. By following the same reasoning as in Sec. I-D, the square Mahalanobis distance D_M^2 can be approximated by $D_M^2(\psi(x);\{\psi_\mu^{[j]}, (C^{[j]})\}) = n_j\tilde{\tilde{\mathbf{k}}}_x^{[j]\top}(\tilde{\tilde{K}}^{[j]})^{-1}\tilde{\tilde{\mathbf{k}}}_x^{[j]}$, which leads to decision functions identical to (22) and (23).

APPENDIX II
DISSIMILARITY DATA USED IN EXPERIMENTS

A collection of dissimilarity data sets is used in our study. They are briefly characterized below:

- *Mucosa* data consist of autofluorescence spectra acquired from healthy and diseased mucosa in the oral cavity by using the excitation wavelength of 365nm; see [35] for details. There are 857/112 normalized spectra representing healthy and diseased tissues, respectively. First-order Gaussian-smoothed derivatives of the spectra are compared by the use of city block distances (smoothing based $\sigma = 3$ samples).
- *Heart* data is a two-class vectorial data set and consists of 303 examples and 13 continuous and categorical features. It comes from the Machine Learning Repository [36] for which the Gower distance was computed as proposed in [37] to

handle different types of features. It is isometric to Euclidean distance.

- *Poly-H* and *Poly-MH* data consist of two classes, 2000 examples each, of randomly generated convex quadrilaterals and irregular heptagons. The polygons are first normalized and then the Hausdorff and modified Hausdorff distances are computed between their vertices [38].
- *News-COR* is a small part of the 20 Newsgroups data, based on four classes: the 'comp.*', 'rec.*', 'sci.*' and 'talk.*' newsgroups. Each message is described by occurrence vectors in a 100-dimensional space. The dissimilarity data is derived based on the non-metric correlation-based measure $d(\mathbf{x},\mathbf{x}') = \frac{1}{2}(1 - \frac{\mathbf{x}^T\mathbf{x}'}{||\mathbf{x}||^2+||\mathbf{x}'||^2-2\mathbf{x}^T\mathbf{x}'})$. The Newsgroups data is available from `http://www.cs.toronto.edu/˜roweis/data.html`.
- *ProDom* is a subset of 2604 protein domain sequences from the ProDom set [39]. We use the four-class problem defined by pairwise structural alignments computed by Roth [40]. The asymmetric similarities are averaged out.
- *Files* data describes five classes of different types of computer files: '*.cc','*.tex','*.bib','*.ps' and '*.pdf'. The asymmetric compression distance, approximating the Kolmogorov complexity, is derived between the files based on the `gzip`-compressor [41]. What is interesting here is that the self-dissimilarity, $d(x,x)$ is usually non-zero.
- *Cat-cortex* dissimilarity data describe the connection strengths between 65 cortical areas of a cat [42]. Four cortex functions are distinguished. The dissimilarity values are measured on the ordinal scale and take the values of $1, 2, 3$ and 4.
- *Protein* data compare the protein sequences based on the concept of an evolutionary distance. The proteins are originally assigned to four classes of globins [43], [44].
- *Pen-angle* and *Pen-dist* refer to pen-based handwritten digit data from the UCI Machine Learning Repository, `http://www.ics.uci.edu/ mlearn/MLRepository.html`. Each digit is represented by a string of 2D vectors on a contour. The distance between the strings is an edit distance with fixed insertion and deletion costs. Two different substitution costs, angle and Euclidean distance between the consecutive vectors lead to two different dissimilarity data. Here, only a part of the pen-digits data is used. See also [45].
- *Zongker* digit data describe ten digit classes, each of 200 examples. Digits are represented by binary images. An asymmetric similarity s_{ij} based on deformable template matching is used to compare the digit shapes [46]. The similarities are averaged out.
- *Chicken15*, *Chicken29* dissimilarity data sets are derived from the five-class Chicken Pieces Silhouettes data [47]. The edges are first approximated by straight line segments of a fixed length $L = 15$ (*Chicken15*) and $L = 29$ (*Chicken29*) and then angle sequences are derived. These are compared by edit distances with fixed insertion and deletion costs equal to 45 degrees and a substitution cost of the absolute difference between the angles. The distances are asymmetric and made symmetric by averaging. The data sets are available from `http://www.iam.unibe.ch/fki/databases`.
- *Chromo-DIF* and *Chromo-ABS* dissimilarity data sets are derived from the 21-class Copenhagen Chromosome Database [48]. The chromosomes are described by strings derived from difference-coded six-level nonlinear profiles. Edit distances are used to derive the dissimilarity sets. Cost functions are defined based on either difference codes (*Chromo-DIF*) or absolute codes (*Chromo-ABS*). The data sets are available from `http://www.iam.unibe.ch/fki/databases`.

References

[35] M. Skurichina and R. Duin, "Combining different normalizations in lesion diagnostics," in *Artificial Neural Networks and Information Processing*, 2003, pp. 227–230.

[36] S. Hettich, C. Blake, and C. Merz, "UCI repository of Machine Learning databases," 1998, http://www.ics.uci.edu/ mlearn/MLRepository.html.

[37] J. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, pp. 25–33, 1971.

[38] M. Dubuisson and A. Jain, "Modified Hausdorff distance for object matching," in *International Conference on Pattern Recognition*, vol. 1, 1994, pp. 566–568.

[39] F. Corpet, F. Servant, J. Gouzy, and D. Kahn, "ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons," *Nucleid Acids Research*, vol. 28, pp. 267–269, 2000.

[40] V. Roth, J. Laub, J. Buhmann, and K.-R. Müller, "Going metric: Denoising pairwise data," in *Advances in Neural Information Processing Systems*, 2003, pp. 841–856.

[41] J. Handl and J. Knowles, "Multiobjective clustering around medoids," in *Operator Theory: Advances and Applications*. IEEE Press, 2005, pp. 550–557.

[42] J. Scannell, C. Blakemore, and M. Young, "Analysis of connectivity in the cat cerebral cortex," *Journal of Neuroscience*, vol. 15, pp. 1463–1483, 1995.

[43] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer, "Classification on pairwise proximity data," in *Advances in Neural Information Processing Systems 11*. MIT Press, 1999, pp. 438–444.

[44] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, and B. Williamson, "Classification on proximity data with LP–machines," in *International Conference on Artificial Neural Networks*, 1999, pp. 304–309.

[45] E. Pękalska, R. Duin, S. Günter, and H. Bunke, "On not making dissimilarities Euclidean," in *Joint IAPR International Workshops on SSPR and SPR*, 2004, pp. 1145–1154.

[46] A. Jain and D. Zongker, "Representation and recognition of handwritten digits using deformable templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1386–1391, 1997.

[47] G. Andreu, A. Crespo, and J. Valiente, "Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition," in *International Conference on Neural Networks*, vol. 2, 1997, pp. 1341–1346.

[48] C. Lundsteen, J. Phillip, and E. Granum, "Quantitative analysis of 6985 digitized trypsin g-banded human metaphase chromosomes," *Clinical Genetics*, vol. 18, p. 355–370, 1980.

Preprints
"Angewandte Mathematik und Informatik"

04/06 - N T. Kösters, F. Wübbeling: Efficient Implementation and Evaluation of Listmode-OSEM based algorithms in PET on shared memory machines

05/06 - I Chr. Jansen, F. Steinicke, J. Vahrenhold, B. Schwald, K. Hinrichs: Enhancing Stereo Tracking via Adapted Point-based Targets

06/06 - S G. Alsmeyer, A. Iksanov, U. Rösler: On distributional properties of perpetuities

01/07 - S D. Völker: Schadenreservierung im Licht Stochastischer Prozesse

02/07 - S G. Alsmeyer, M. Meiners: A Stochastic Maximin Fixed Point Equation Related to Game-Tree Evaluation

03/07 - I T. Ropinski, J. Meyer-Spradow, S. Diepenbrock, J. Mensmann, K. Hinrichs: Interactive Volume Rendering Supporting Global Illumination Phenomena

04/07 - I J. Müller-Iden, S. Gorlatch, T. Schröter, S. Fischer: Entity Density Scalability of Multiplayer Online Games via Replication-based Parallelization: A Case Study of Quake 2

05/07 - I F. Steinicke, G. Bruder, K. Hinrichs: Navigation Metaphors for Head-Mounted Display Environments

06/07 - N P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, O. Sander: A Generic Grid Interface for Parallel and Adaptive Scientific Computing, Part I: Abstract Framework

07/07 - N P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, O. Sander: A Generic Grid Interface for Parallel and Adaptive Scientific Computing, Part II: Implementation and Tests in DUNE

08/07 - I F. Steinicke, J. Mensmann, K. Hinrichs, K. Rothaus, J. de Buhr, A. Krüger: Real-Time Rendering of Weather-Related Phenomena in Digital 3D Urban Models

09/07 - N B. Haasdonk, M. Ohlberger, G. Rozza: A Reduced Basis Method for Evolution Schemes with Parameter-Dependent Explicit Operators

10/07 - S G. Alsmeyer: Minimal Position and Critical Martingale Convergence in Branching Random Walks

01/08 - N P. Henning, M. Ohlberger: The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains

02/08 - S G. Alsmeyer, A. Iksanov: A log-type moment result for perpetuities and its application to martingales in supercritical branching random walks

03/08 - S G. Alsmeyer, M. Meiners: On a Min-Type Stochastic Fixed-Point Equation Related to the Smoothing Transformation

04/08 - S G. Alsmeyer, M. Meiners: A Note on the Transience of Critical Branching Random Walks on the Line

05/08 - S G. Alsmeyer, G. Hölker: Asymptotic Behavior of Ultimately Contractive Iterated Lipschitz Functions

06/08 - N E. Pekalska, B. Haasdonk: Kernel Quadratic Discriminant Analysis with Positive Definite and Indefinite Kernels