Kernel Fisher's Discriminant Analysis
in Gaussian Reproducing Kernel Hilbert Space[1]

Su-Yun Huang[2]   syhuang@stat.sinica.edu.tw

Institute of Statistical Science

Academia Sinica, Taipei 11529, Taiwan, R.O.C.

Chii-Ruey Hwang   crhwang@sinica.edu.tw

Institute of Mathematics

Academia Sinica, Taipei 11529, Taiwan, R.O.C.

Miao-Hsiang Lin   miao@stat.sinica.edu.tw

Institute of Statistical Science

Academia Sinica, Taipei 11529, Taiwan, R.O.C.

Feb. 1, 2005

Abstract

Kernel Fisher's linear discriminant analysis (KFLDA) has been proposed for non-linear binary classification (Mika, Rätsch, Weston, Schölkopf and Müller, 1999, Baudat and Anouar, 2000). It is a hybrid method of the classical Fisher's linear discriminant analysis and a kernel machine. Experimental results (e.g., Schölkopf and Smola, 2002) have shown that the KFLDA performs slightly better in terms of prediction error than the popular support vector machines and is a strong competitor to the latter. However, there is very limited statistical justification of this method. In this article we provide a fundamental study for it in the framework of a Gaussian reproducing kernel Hilbert space (RKHS) and give an extension of the KFLDA to a quadratic generalization, called kernel Fisher's quadratic discriminant analysis (KFQDA).

In our approach each data point is mapped to a function in the kernel associated RKHS. Next the problem of Fisher's discriminant is solved in this new kernel data space. This kernelized Fisher's approach can be regarded, from the original data space viewpoint, as a nonparametric approach to classification since it adopts kernels as basis functions and its decision boundary is modeled via kernel mixture. Still it has the computational advantage of keeping the training process analogous to a parametric method. We show that the kernel Fisher's discriminant can be obtained as a maximum likelihood method and a Bayes classifier under Gaussian assumption in the associated RKHS. We also demonstrate that our kernel transformations can drastically draw the data distribution to better Gaussian. Theorems are given to show that, under suitable conditions, most projections of data represented via kernel functions in a RKHS are approximately Gaussian, which justifies the Gaussian assumption in the RKHS.

Key words and phrases: Bayes classifier, kernel Fisher's discriminant analysis, Gaussian reproducing kernel Hilbert space, maximum likelihood method, projection pursuit.

---

[1]Running title: Kernel Fisher's Discriminant Analysis in Gaussian RKHS.

[2]Corresponding author.

1   Introduction

The aim of discriminant analysis is to classify an object into one of k given groups based on training data consisting of $\{(x_j, y_j)\}_{j=1}^n$, where $x_j \in \mathcal{X} \subset R^p$ is a p-variate input measurement and $y_j \in \{1, \ldots, k\}$ indicates the corresponding group membership. The classical Fisher's linear discriminant analysis (FLDA) is a commonly used and time-honored tool for multiple classification because of its simplicity and probabilistic outputs. With $k \leq p + 1$, FLDA finds $k - 1$ canonical variates that are optimal (in a certain sense) for separating the groups, and the FLDA's decision boundaries are linear in these canonical variates. Often such a linear formulation for decision rule is not adequate, thus, quadratic decision boundaries are called for. But still there are commonly seen cases which need more general nonlinear decision rule. Motivated from the active development of statistical learning theory (Vapnik, 1998; Hastie, Tibshirani and Friedman, 2001) and the popular and successful usage of various kernel machines (Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002), there has emerged a hybrid approach which combines the idea of feature map in SVM with the classical Fisher's linear discriminant approach. Its usage can be traced back to Mika, Rätsch, Weston, Schölkopf and Müller (1999), and Baudat and Anouar (2000). Later it was also studied by Mika, Rätsch and Müller (2001), Van Gestel, Suykens and De Brabanter (2001), Mika, Smola and Schölkopf (2001), Xu, Zhang and Li (2001), and Mika (2002). However, there is very limited statistical justification of this method despite its successful performance in classification. In this article we provide a fundamental study for it in the framework of Gaussian reproducing kernel Hilbert spaces (RKHS).

Mika, Rätsch, Weston, Schölkopf and Müller (1999), Baudat and Anouar (2000) and others have used the term "KFDA" for the hybrid method of the FLDA and a kernel machine. However, throughout this work we will use the term KFLDA to indicate such a kernel-Fisher-linear approach, as later we will extend the KFLDA to a quadratic generalization, called kernel Fisher's quadratic discriminant analysis (KFQDA). The abbreviation KFDA is then used for kernel Fisher's approach in general including KFLDA and KFQDA. Kernels used in KFDA are positive definite kernels (also known as reproducing kernels, Aronszajn, 1950). In all the articles mentioned above, the KFLDA is based on performing the FLDA in a kernel-spectrum-based feature space. In this article, our KFDA will be based on performing Fisher's procedure directly right in the associated RKHS rather than in the kernel-spectrum-based feature space. Our KFDA is a two-stage procedure. The first stage is to embed the data space $\mathcal{X}$ into an infinite dimensional RKHS, denoted by $\mathcal{H}_\kappa$, via a kernel function κ. The second stage is to carry out Fisher's procedure in this new data space $\mathcal{H}_\kappa$. As the second stage is exactly the same as the classical Fisher's approach, existing software codes are ready for use, for instance, SAS (proc discrim), Matlab (classify), Splus (discr) and R (lda, qda). The only extra effort involved is in the first stage to prepare the original data by a new data representation, namely the kernel data representation. In other words, the working procedure for KFDA is rather simple as the training process is operated via the notion of the parametric FDA. On the other hand, the underlying model behind the KFDA is sophisticated and versatile, which is due to the fact that decision boundaries are built from kernel mixtures.

Our kernelized procedure of KFDA leads to generalize the classical FDA so that flexible nonlinear features in training inputs can be better explored and included into the decision rule. Data, which are embedded into an infinite dimensional RKHS, become sparse and turn to be better elliptically symmetrically distributed and then can be better separated by simple parametric boundaries (herein linear or quadratic in the associated RKHS). The classical FDA finds an optimal (in a certain sense) low dimensional subspace and then the discrimination is done in this low-dimensional subspace. It can be derived as a maximum likelihood method as well as a Bayes classifier under Gaussian assumption on $\mathcal{X}$. Parallel to the classical theory the KFDA proposed in this article extends the maximum likelihood method and Bayes classification to their kernel generalization under Gaussian Hilbert space assumption. Theorems will be given to justify such a Gaussian Hilbert space assumption. We show that under suitable conditions most low-dimensional projections of functional data are approximately Gaussian. This extends part of Diaconis and Freedman's (1984) results to functional data. Readers are referred to their Theorems 1.1 and 1.2 on limiting distributions for low-dimensional projections of high-dimensional data, and Example 3.1 for iid coordinates.

The rest of the article is organized as follows. In Section 2 we give a brief review of KFLDA in the kernel-spectrum-based feature space. In Section 3 we introduce an alternative feature map and demonstrate by empirical data study that such an alternative map can drastically improve the Gaussianity of data distribution. In Section 4 we introduce the KFDA methodology and its implementation. In Section 5 we give an experimental study using a real data set, a synthetic data set and some benchmark data sets. In Section 6 we give a fundamental study on the theory of KFDA. All proofs are in the Appendix.

2 Review: KFLDA in feature space

This section gives a very brief review of KFLDA. The KFLDA procedure in Mika *et al.* (1999) and Baudat and Anouar (2000) was formulated in a kernel-spectrum-based feature space. For a given positive definite kernel and its spectrum:

$$\kappa(x, u) = \sum_{q=1}^{d} \lambda_q \psi_q(x) \psi_q(u), \tag{1}$$

the main idea of the KFLDA is first to map the data in the input space $\mathcal{X} \subset R^p$ into the spectrum-based feature space, denoted by $\mathcal{Z} \subset R^d$ with $d \leq \infty$, via a transformation

$$x \rightarrow z(x) = (\sqrt{\lambda_1}\psi_1(x), \ldots, \sqrt{\lambda_d}\psi_d(x))' = \Psi(x).$$

Next the classical Fisher's procedure is operated on the transformed data in the feature space $\mathcal{Z}$. For a binary classification, the KFLDA finds the discriminant function of the form

$$d(z) = w'z + b = \sum_{q=1}^{d} w_q \sqrt{\lambda_q}\, \psi_q(x) + b, \tag{2}$$

where w is the canonical variate that maximizes the so-called Rayleigh coefficient in the feature space $\mathcal{Z}$

$$J_{KFLDA}(w) \equiv \frac{w'S_b w}{w'S_w w + rA(w)},$$

where $S_b = (\bar{z}_1 - \bar{z}_2)(\bar{z}_1 - \bar{z}_2)'$ and $S_w = \sum_{j=1}^{n} \{z_j z_j' - (n_1 \bar{z}_1 \bar{z}_1' + n_2 \bar{z}_2 \bar{z}_2')\}/(n-2)$ are respectively the between- and within-group sample covariances in $\mathcal{Z}$, $\bar{z}_1$, $\bar{z}_2$, n_1, n_2 are group means and group sizes in $\mathcal{Z}$, and $A(\alpha)$ is a penalty functional on w. In practice, the kernel function κ is defined directly without explicit expression for its spectrum Ψ. Thus, there is no way that S_b and S_w are explicitly known. However, as the inner product in feature space can be represented via the following kernel value:

$$\langle z(x), z(u) \rangle_{\mathcal{Z}} = \sum_{q=1}^{d} \lambda_q \psi_q(x)\psi_q(u) = \kappa(x, u), \tag{3}$$

it allows us to work directly on the kernel values without knowing the spectrum-based transformation $\Psi : \mathcal{X} \to \mathcal{Z}$, nor the associated sample means and covariances. One can prove that (see, for instance, Mika *et al.*, 1999) the solution w can be expanded as

$$w = \sum_{j=1}^{n} \alpha_j \Psi(x_j) = Z'\alpha, \tag{4}$$

where $Z = (\Psi(x_1), \ldots, \Psi(x_n))'$ is the transformed input data matrix in $\mathcal{Z}$. The discriminant function can be re-formulated as

$$d(x) = \sum_{j=1}^{n} \alpha_j \kappa(x_j, x) + b. \tag{5}$$

The coefficients α_j's can then be obtained as the solution to the following optimization problem

$$\arg \max_{\alpha \in R^n} J_{RKFLDA}(\alpha) \equiv \arg \max_{\alpha \in R^n} \frac{\alpha' M_b \alpha}{\alpha' M_w \alpha + rA(\alpha)}, \tag{6}$$

where $M_b = (\bar{k}_1 - \bar{k}_2)(\bar{k}_1 - \bar{k}_2)'$, $M_w = (K^2 - \sum_{i=1}^{2} n_i \bar{k}_i \bar{k}_i')/(n-2)$, $K = [\kappa(x_j, x_{j'})]_{n \times n}$, $\bar{k}_i = n_i^{-1} \sum_{j \in I_i} K_j$, K_j is the j-th column of K and I_i the index set for group i. Since the matrix M_w is at most of rank $n-1$, a penalty functional $A(\alpha)$ is added to overcome the numerical problem caused by singular within-group covariance M_w. With α being solved from (6), the intercept b of the discriminant hyperplane (5) is determined by setting the hyperplane to pass through the mid point of the two group means, i.e., $b = -\alpha'(\bar{k}_1 + \bar{k}_2)/2$.

3 Aronszajn kernel map and data normality

The kernel trick in (3) of turning inner products in $\mathcal{Z}$ into kernel values allows us to carry out the KFLDA in the spectrum-based feature space without explicitly knowing the spectrum itself. However, without knowing the spectrum transformation $z = \Psi(x)$ there is no way of expressing a quadratic form $\langle z, Tz \rangle_{\mathcal{Z}}$ in terms of kernel values, which hinters the extension

of KFLDA to a kernel-quadratic classifier. Here, instead of the spectrum-based feature map $\Psi(x)$, we introduce an alternative way of data representation (7). Our representation embeds the data space $\mathcal{X}$ directly into the kernel associated RKHS and allows an arbitrary quadratic form to be expressed in terms of kernel values to admit a hybrid of FQDA with a kernel machine. Example 3 in Section 5 gives a view that a kernel-linear procedure might not be as adequate or efficient as a kernel-quadratic procedure. Also with the new kernel map (7) the KFDA can be reformulated so that statistical properties, like maximum likelihood ratio and Bayes classification, can be naturally developed in this framework.

In this section we focus mainly on introducing this alternative data representation and its effect on improving data normality. The classical FDA is good for data with predictors having approximately normal distribution or at least having approximately elliptically symmetric distribution. The normality or elliptical symmetry is a very restrictive condition on data distribution. A way to improve the data normality is to embed the data space $\mathcal{X}$ into a very high dimensional (often infinite dimensional) space, called the embedding sample space. Then project the embedding sample space into a low dimensional subspace. Data normality can be drastically improved if handled in this way. Below we introduce an embedding of bringing data to a high dimensional space via kernels.

Kernels used throughout the article are positive definite kernels (also known as reproducing kernels). See Aronszajn (1950) for theory of reproducing kernel and reproducing kernel Hilbert space. Given a positive definite kernel κ on $\mathcal{X} \times \mathcal{X}$, we are going to associate with it a RKHS.

Definition 1 (Reproducing kernel Hilbert space) *A RKHS is a Hilbert space of real-valued functions on $\mathcal{X}$ satisfying the property that, all the evaluation functionals are bounded linear functionals.*

To every RKHS there corresponds a unique positive-definite kernel κ such that $\langle f, \kappa(x, \cdot) \rangle = f(x)$ for all f in this RKHS. We say that this RKHS admits the kernel κ. Conversely, given a positive-definite kernel κ on $\mathcal{X} \times \mathcal{X}$ there exists a unique Hilbert space admitting this kernel. We denote this Hilbert space by $\mathcal{H}_\kappa$.

Consider a transformation $\Gamma : \mathcal{X} \to \mathcal{H}_\kappa$ given by

$$x \to \Gamma(x) = \kappa(x, \cdot). \tag{7}$$

The data space $\mathcal{X}$ is embedded into a new data space, i.e., the embedding sample space $\mathcal{H}_\kappa$, via the transformation Γ. Each input point $x \in \mathcal{X}$ is mapped to a function $\kappa(x, \cdot) \in \mathcal{H}_\kappa$, which is called Aronszajn map in Hein and Bousquet (2004). Their article gives a survey of results in the mathematical literature on positive definite kernels and associated structures potentially relevant for machine learning. If we place a probability distribution P on $\mathcal{X}$, correspondingly there is an induced probability measure on $\mathcal{H}_\kappa$. In Example 1 below we show by a synthetic data experiment that the Aronszajn map Γ can draw the data distribution closer to Gaussian. We first introduce the Gaussian measure on an arbitrary Hilbert space $\mathcal{H}$. One way to define Gaussian measure on a Hilbert space is through joint normal distributions of inner products between the random element and an arbitrarily selected finite "coordinate system" $\{f_1, \ldots, f_m\}$. The definition goes as follows.

Definition 2 (Gaussian measure on a Hilbert space) *Let $\mathcal{H}$ be an arbitrary real separable Hilbert space. A probability measure $P_{\mathcal{H}}$ defined on $\mathcal{H}$ is said to be Gaussian, if for any m and any $\{f_1, \ldots, f_m \in \mathcal{H}\}$, the joint distribution of*

$$\langle f_1, h \rangle_{\mathcal{H}}, \ldots, \langle f_m, h \rangle_{\mathcal{H}}$$

is normal, where h is the random element in $\mathcal{H}$ having the probability measure $P_{\mathcal{H}}$. By Riesz Representation Theorem, $\{f_1, \ldots, f_m \in \mathcal{H}\}$ can be replaced by $\{\ell_1, \ldots, \ell_m\}$, where ℓ_j's are bounded linear functionals on $\mathcal{H}$.

See Grenander (1963), Vakhania, Tarieladze and Chobanyan (1987), and Janson (1997) for Gaussian measures on Hilbert spaces.

Example 1 Draw 10000 random samples from an exponential distribution with probability density function $p(x) = \exp(-x)$. Arrange them in a 200×50 matrix, denoted by X. The data matrix X is used to represent a random sample made up with 200 random vectors in R^{50}. Each row of X is a 50-dimensional data point and there are 200 data points in total. Each column of X represents a projection of this random sample along a one-dimensional coordinate axis. Render all these 50 coordinate columns to normality checks using (1) the Kolmogorov-Smirnov test and (2) the normal probability plot. The average p-value and its standard error of these 50 Kolmogorov-Smirnov tests are:

$$\text{average} \ \ p - \text{value} = 2.853 \times 10^{-4}, \ \ \text{standard error} = 5.633 \times 10^{-4}.$$

We also plot the normal probability plots for the best 4 out of 50 coordinate columns closest to Gaussian (Figure 1), the median 4 coordinate columns (Figure 2), and the worst 4 coordinate columns farthest from Gaussian (Figure 3). Associated p-values for the best four, median four and worst four cases are reported below:

$$
\begin{aligned}
best \ 4 \ \ &: \ \ 0.0025 \ \ 0.0023 \ \ 0.0018 \ \ 0.0010 \\
median \ 4 \ \ &: \ \ 0.0000 \ \ 0.0000 \ \ 0.0000 \ \ 0.0000 \\
worst \ 4 \ \ &: \ \ 0.0000 \ \ 0.0000 \ \ 0.0000 \ \ 0.0000
\end{aligned}
$$

From these p-values and normal probability plots in Figures 1-3, it is clear that this random sample of size 200 from a 50-dimensional exponential distribution is far from being Gaussian.

To improve the Gaussianity of data distribution, we transform the data by the Aronszajn kernel map Γ in (7). We have tried out using both the Gaussian kernel and the Epanechnikov kernel. Via the kernel map Γ, each data point $x_j \in R^{50}$ is given a new representation by a function. For the case of Gaussian kernel,

$$x_j \rightarrow \kappa(x_j, x) = \exp(-0.5\|x - x_j\|_2^2/\sigma^2) \ \text{ with } \sigma = 10.$$

For the case of Epanechnikov kernel, let x_k and x_{jk} denote the kth coordinate values of x and x_j respectively, i.e., $x = (x_1, \ldots, x_k, \ldots, x_{50})$ and $x_j = (x_{j1}, \ldots, x_{jk}, \ldots, x_{j50})$, and let

$$x_j \rightarrow \kappa(x_j, x) = \prod_{k=1}^{50} (1 - (x_k - x_{jk})^2/\sigma^2)\mathcal{I}\{|x_k - x_{jk}| < \sigma, \forall k = 1, \ldots, 50\} \ \text{ with } \sigma = 10,$$
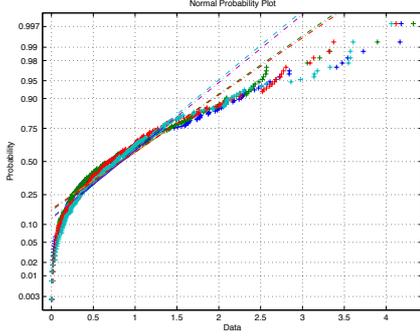
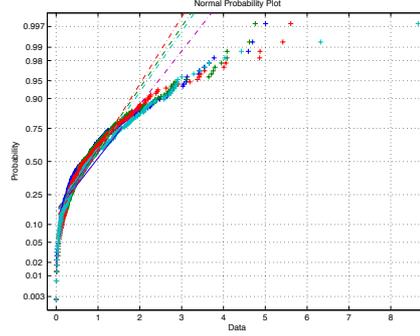Figure 1: Best four, original data.
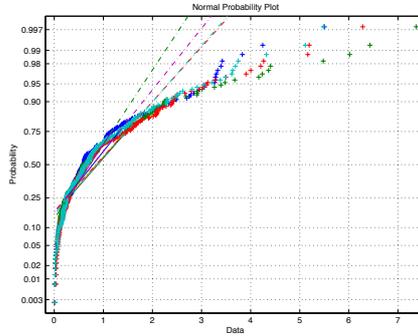

Figure 2: Median four, original data.

Figure 3: Worst four, original data.

where $\mathcal{I}$ is an indicator function. We then check if the empirical distribution of $\{\kappa(x_j, \cdot)\}_{j=1}^{200}$ resembles a random sample from a Gaussian measure on $\mathcal{H}_\kappa$. Based on Definition 2, we choose a coordinate system $\{\ell_j\}_{j=1}^{200}$ to discretize $\kappa(x_j, \cdot)$. The ℓ_j's are taken to be evaluation functionals at x_j's, i.e., $\ell_j(h) = h(x_j)$ for $h \in \mathcal{H}_\kappa$. This particular choice of coordinate system leads to a resulting input kernel matrix, denoted by K, given by $[\kappa(x_j, x_{j'})]_{200 \times 200}$. This kernel data matrix is same as that in the conventional SVM and in KFLDA of Mika *et al.* (1999). One may choose to use a different coordinate system of different size to discretize $\kappa(x_j, \cdot)$. The kth column of K comes from projecting the kernel data $\{\kappa(x_j, \cdot)\}_{j=1}^{200}$ along the functional direction determined by the evaluation functional ℓ_k, which is $\kappa(x_k, \cdot)$. That is, columns in K are the result of one-dimensional projections of $\{\kappa(x_j, \cdot)\}_{j=1}^{200}$. These columns are then rendered to normality checks. Results show that kernel-transformed data are much closer to a Gaussian distribution. Listed below are summarized p-values for the best four, median four and worst four out of 200 coordinate columns of transformed data. For the case of Gaussian kernel, we have

$$
\begin{aligned}
best\ 4 &: \quad 0.9943 \ \ 0.9794 \ \ 0.9727 \ \ 0.9634 \\
median\ 4 &: \quad 0.5771 \ \ 0.5769 \ \ 0.5702 \ \ 0.5658
\end{aligned}
$$

$$worst\ 4\ \ :\ \ 0.0975\ \ 0.0854\ \ 0.0288\ \ 0.0202$$
$$average\ \ :\ \ 0.5650\ \ (std\ \ error = 0.2542)$$

For the case of Epanechnikov kernel, we have

$$best\ 4\ \ :\ \ 0.9583\ \ 0.9527\ \ 0.9134\ \ 0.9063$$
$$median\ 4\ \ :\ \ 0.3301\ \ 0.3291\ \ 0.3123\ \ 0.3122$$
$$worst\ 4\ \ :\ \ 0.0263\ \ 0.0261\ \ 0.0210\ \ 0.0044$$
$$average\ \ :\ \ 0.3662\ \ (std\ \ error = 0.2441)$$

Normal probability plots are presented in Figures 4-6 for Gaussian-kernel transformed data, and in Figures 7-9 for Epanechnikov-kernel transformed data. It is clear that kernel (Gaussian or Epanechnikov alike) transformed data are much better Gaussian than the untransformed data.
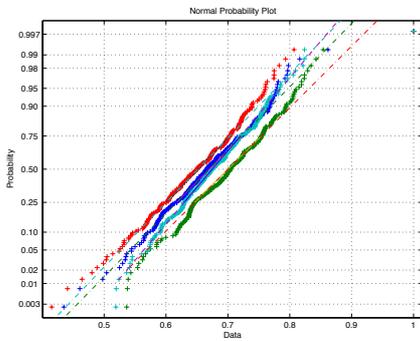

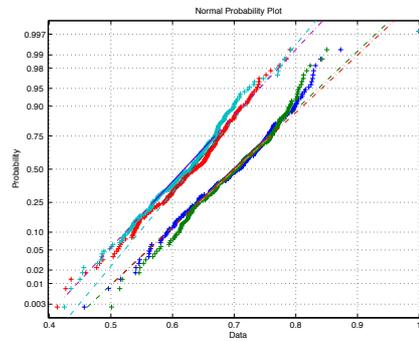Figure 4: Best four columns, Gaussian.

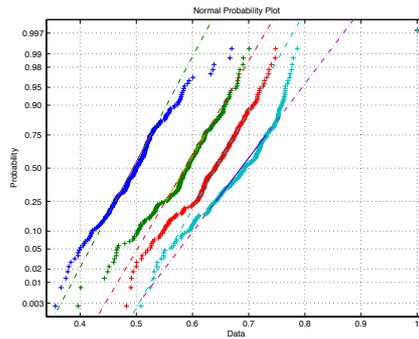Figure 5: Median four columns, Gaussian.
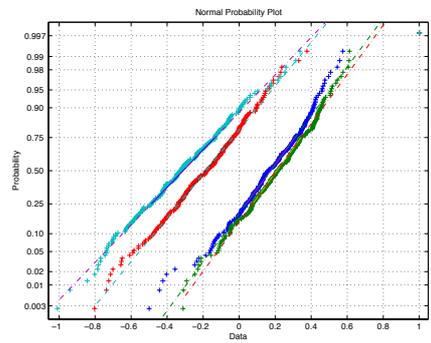

Figure 6: Worst four columns, Gaussian.
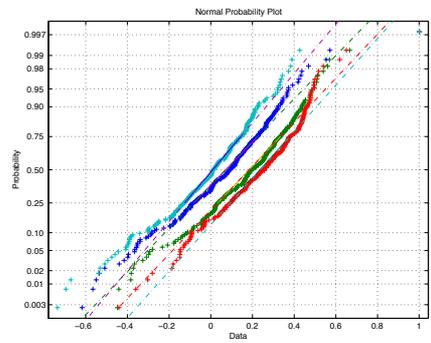
Figure 7: Best four columns, Epanechnikov.
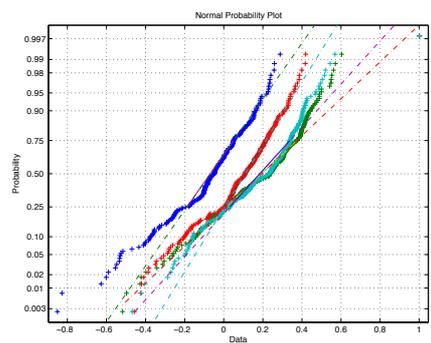

Figure 8: Median four columns, Epanechnikov.

Figure 9: Worst four columns, Epanechnikov.

4   Methodology

4.1   KFDA in RKHS

Let $\pi_1, \ldots, \pi_k$ denote the underlying populations. Let I_i be the index set of training sample from π_i and $I = \cup_{i=1}^{k} I_i$ be the index set for the entire data. The training data can then be partitioned as $\cup_{i=1}^{k} \{(x_j, y_j)\}_{j \in I_i}$. Let $n_i = |I_i|$, the size of I_i, and $n = |I|$, the size of I. Let $X = (x_1, \cdots, x_n)'$ be an $n \times p$ matrix consisting of training inputs. We use the notation

$$[X, y] \quad \text{for original training data,}$$

which is an $n \times (p+1)$ matrix including training inputs and their corresponding group labels. Suppose that π_i, has probability density function $f_i(x)$, $i = 1, \ldots, k$. Also assume that the prior probability of an observation coming from π_i is q_i. Then the conditional probability of a given input measurement x coming from population π_i is

$$prob(\pi_i | x) = \frac{q_i f_i(x)}{q_1 f_1(x) + \cdots + q_k f_k(x)}, \quad i = 1, \ldots, k. \tag{8}$$

An equivalent expression for $prob(\pi_i | x)$ is

$$prob(\pi_i | x) = \frac{q_i f_i(x)/q_1 f_1(x)}{1 + q_2 f_2(x)/q_1 f_1(x) + \cdots + q_k f_k(x)/q_1 f_1(x)}, \quad i = 1, \ldots, k. \tag{9}$$

Thus, we assign π_i for x if $prob(\pi_i | x)$ is the maximum. By this conditional probability approach it is sufficient to train as many as $k - 1$ binary classifiers of π_i against π_1 for $i = 2, \ldots, k$. If $\pi_1, \ldots, \pi_k$ can be approximated by Gaussian distributions with a common covariance, then the decision boundaries are linear. If these Gaussian populations do not share a common covariance, then (8), or equivalently (9), leads to quadratic decision boundaries. However, if $\pi_1, \ldots, \pi_k$ are not Gaussian in $\mathcal{X}$, but can be approximated by Gaussian distributions after being embedded via Γ into $\mathcal{H}_\kappa$. The problem then becomes to solve the Fisher's multiple classification in the RKHS $\mathcal{H}_\kappa$. (Detailed theoretic derivations are shown in Section 6.) The KFDA solves a classification problem in two steps. First, it computes the kernel data arranged in a matrix K with (j, j')th entry given by $K_{jj'} = \kappa(x_j, x_{j'})$. Secondly, it performs the classical Fisher's procedure on

$$[K, y] \quad \text{kernel training data,} \tag{10}$$

in the same way as on the original data $[X, y]$. Each row of K, say jth row, is treated as a training input with a corresponding group label y_j in the same row. A test input x is transformed via Γ to a function $\kappa(x, \cdot)$ with discretization $K_x = (\kappa(x, x_1), \ldots, \kappa(x, x_n))'$. The KFLDA decision functions are given by

$$\log(q_i f_i(x)/q_1 f_1(x)) = \{K_x - (\bar{k}_i + \bar{k}_1)/2\}' M^{-1}(\bar{k}_i - \bar{k}_1) + \rho_i \tag{11}$$

where $\rho_i = \log(q_i/q_1)$, M is the pooled sample covariance of kernel training inputs K, and $\bar{k}_i$ is the sample mean of kernel training inputs from index set I_i. The test input x is assigned

to π_i, if its log-likelihood ratio is the maximum. When the transformed data do not share a common covariance, then we use the following KFQDA decision functions

$$
\begin{aligned}
&\log(q_i f_i(x)/q_1 f_1(x)) \\
&= \frac{1}{2}\{(K_x - \bar{k}_1)' M_1^{-1} (K_x - \bar{k}_1) - (K_x - \bar{k}_i)' M_i^{-1} (K_x - \bar{k}_i)\} + \rho_i,
\end{aligned} \tag{12}
$$

where $\rho_i = \log(q_i/q_1) + \frac{1}{2}\log(|M_1|/|M_i|)$, M_i's are within group sample covariances and $|M_i|$'s are matrix determinants.

Remark 1 (kernel Mahalanobis) *An alternative choice of quadratic decision functions is to remove the term $\frac{1}{2}\log|M_1|/|M_i|$ from ρ_i. In this case, (12) is a kernel Mahalanobis distance based criterion.*

Remark 2 (discretization) *The kernel data in (10) are based on a particular way of discretization as explained in Example 1. There are other alternative discretization methods. For instance, one may consider to use a discretizing coordinate system consisting of $\{\ell_{u_j}\}_{j=1}^m$, where ℓ_{u_j}'s are evaluation functionals at points u_j, i.e., $\ell_{u_j}(h) = h(u_j)$ for $u_1, \ldots, u_m \in \mathcal{X}$ and $h \in \mathcal{H}_\kappa$. This discretization is also known as empirical kernel map (Schölkopf and Smola, 2002). Later in Section 4.2 we will introduce another discretization: the random subset approach (Lee and Mangasarian, 2001).*

Remark 3 (regularization) *The discretization used in (10) results in singular sample covariance matrices M and M_i. Some kind of regularization is necessary. In Section 4.2 we will introduce 3 regularization methods: the principal components approach, the random subset approach and the ridge approach (Mika et al., 1999).*

Remark 4 (discriminant analysis by Gaussian mixtures) *With some derivation the decision functions (11) can be obtained as*

$$
d(x) = \sum_{j=1}^n \alpha_j \kappa(x_j, x) + b, \tag{13}
$$

which is a mixture of kernels. If $\kappa(\cdot, \cdot)$ is a Gaussian kernel, the KFLDA is equivalent to a discriminant analysis by Gaussian mixture. The Gaussian mixture approach is not new in statistical and pattern recognition literature, see, for instance, Hastie and Tibshirani (1996), Taxt, Hjort and Eikvil (1991). However, the KFLDA as a Gaussian mixture has two main attractive features over other Gaussian mixture approaches. One is that, the KFLDA uses a nonparametric modelling scheme, but its implementation algorithm uses parametric notion and the training process is exactly the same as the classical version. The other is that, the KFDA possesses some statistical optimality properties, which are shown later in Section 6.

4.2 Implementation and software

The KFDA approach taken here is to embed the data into a RKHS and to solve the classification problem in this new data space using the notion of FDA. The FDA step here is

exactly the same as the classical version. Thus, codes from mathematical and statistical packages are ready for use, for instance, Matlab (classify), Splus (discr), R (lda, qda) and SAS (proc discrim). The only extra work required is to prepare the training and test inputs in kernel form.

The KFLDA uses a saturated model representation (13) for decision functions built from bases $\{\kappa(x_j, \cdot)\}_{j=1}^{n}$, and there are as many parameters α_j as data points. The KFQDA uses an even over-saturated representation. Fitting data to a saturated or an over-saturated model is deemed to be unstable (solutions not unique) and causes prediction error to rise due to over-fitting, unless some sort of regularization is applied. We introduce three different approaches to reduce model complexity and stabilize computation.

Random subset approach. The random subset (RS) approach was originally proposed by Lee and Mangasarian (2001) to overcome computational problems in SVM for massive data. Problems confronted in massive data training are: the size of the mathematical programming problem, the dependence of the decision function on most of the full data bases creating unwieldy storage problems, and problems of prediction instability and over-fitting. The random subset approach, though, randomly selects a small portion of basis functions from the full set to generate the decision function, it fits the entire data set to this reduced model. Each candidate basis in the full set has equal chance of being selected. This uniform random subset approach is simple and straightforward without resorting to any search algorithm for optimal bases, and yet effective enough for building up decision boundaries. We call $\{\kappa(x_j, \cdot)\}_{j=1}^{n}$ the full-data bases and $\{\kappa(\tilde{x}_j, \cdot)\}_{j=1}^{\tilde{n}}$ a reduced set of bases, where $\{\tilde{x}_j\}_{j=1}^{\tilde{n}}$ is a random subset of $\{x_j\}_{j=1}^{n}$. Adopting the random subset approach, the KFDA operates on a much thinner kernel matrix $\tilde{K}$ than the full square matrix K, where $\tilde{K} = [\kappa(x_j, \tilde{x}_\ell)]_{n \times \tilde{n}}$ with $\tilde{n} \ll n$. That is, the FDA step is operated on the reduced kernel data $[\tilde{K}, y]$.

The random subset approach is designated for massive data set. Statistical theory for this approach can be found in Huang and Lee (2004). It is suggested in Lee and Mangasarian (2001) that the reduced set size be about 1% to 2% of the full set for massive data. The random subset approach also works well for moderate-sized data set, and it is suggested that the reduced subset size be about 10% to 20% of the full set. The Pen-digits example and the Adult-data example below use the KFLDA-RS approach.

PCA approach. For a small-sized data set, we suggest the PCA approach. The PCA approach finds the first few, say $\tilde{n} \ll n$, principal components of the column space spanned by the kernel training input K. Then project K along these $\tilde{n}$ principal components to get the reduced-column kernel matrix, $\tilde{K}$, an $n \times \tilde{n}$ matrix. Example 2 (Educational placement) and Example 3 (synthetic data) below use the KFLDA-PCA and KFQDA-PCA, respectively.

Ridge approach. The ridge approach is to add an extra term, $rA(\alpha) = r\alpha'A\alpha$, to the sample covariance as shown in (6). Often A is an identity matrix. Sparse greedy approximation algorithm has been developed in Mika, Smola and Schölkopf (2001) to solve the optimization problem (6).

5 Experiments

Schölkopf and Smola (2002, see Table 15.1) showed a summary of comparison between KFLDA, SVM, a single radial basis function classifier, AdaBoost, and regularized AdaBoost over 13 benchmark data sets. They reported that KFLDA performed either equally well or slightly better than the SVM-type classifiers. To enhance knowledge on the performance of KFDA, we include in this article a few more examples with various data characteristics different from Schölkopf and Smola's, and we also include the extension to KFQDA. Example 2 is a real data set from social science, serving as a typical example of highly imbalanced group sizes. Example 3 uses a synthetic data set to show that KFLDA can be inefficient and/or inadequate, and, thus, KFQDA is required. It is also shown that different kernels have little effects on classification results if a right tuning parameter has been used. Examples 4 and 5 are two benchmark data of moderate to massive size and both serve to show that the KFDA is easily applicable to handle massive data sets. In all our examples below, training and test processes are done using codes from Matlab (classify) or SAS (proc discrim). One thing worth mentioning is that the KFDA procedure can easily deal with a multiple classification problem. Making use of the notion of likelihood ratio and conditional probability, extension from binary to multiple classification under KFDA as compared with SVM is straightforward and less computationally intensive. This is due to the fact that the former is to train $k-1$ binary classifiers, while the latter involves possibly controversial voting scheme and much heavier computation. (Multiple FDA is implemented in all the above mentioned statistical/mathematical softwares.) Another advantage of KFDA over SVM is the robustness in classification for highly imbalanced data. Example 2 below is a real data example with group proportions: 6.16%, 86.55% and 7.28%.

Example 2 (Educational placement) Data used in this example is a partial collection from a project commissioned by Taiwan's National Science Council for developing educational indicators to monitor and upgrade Taiwan's elementary and secondary science education (Cheng, et al., 1994, Lin, Huang and Chang, 2004). The five indicators of sci-

Table 1: Description of group labels and discriminant variables

variable type	description	variable values
group labels	group based on GPA	1,2,3
discriminant variables	scores on nature science test	0-32
	hours spent on studying homework	0-12
	students' attitudes toward science	0-5
	teachers' dedication	0-10
	parents' educational levels	1-6

ence education, developed for assessing sixth graders, are the nature science test, the four questionnaire scales on students' interest and attitudes toward science, teachers' dedication, parents' educational levels, and hours spent on studying homework. The information on the sample also includes students' school GPA. The distribution of GPA scores was used

to designate the students' group membership. The designation procedure was to obtain three relatively exclusive groups by separating students in the bottom and top 5% of the distribution of GPA scores from those in the remaining 90%. This percentage is frequently used in the selection of high- and low-ability students. In addition, students with GPAs in the indifference zones, between the 90+% and 95% quantiles, and between the 5+% and 10% quantiles, were excluded from the study. This was done because no agreement could be made on how to identify students falling in the indifference zones (Glass, 1978). Following this procedure, the remaining students were identified, as the three reference groups in compliance with the placement into the remedial-, regular-, and advanced-curriculum programs, respectively. Hence, the students whose group memberships were established were treated as a training data set with a known group membership status. After removing the indifferent zone students, there were 357 left in the training set, 22 are classified as low ability, 309 as medial ability and 26 as high ability.

We have tried out FLDA, KFLDA (PCA-version) and SVM using leave-one-out cross validation on the training data. Gaussian kernels with covariance matrix $\mathrm{diag}(10S_1, \ldots, 10S_5)$ are used in both KFLDA and SVM, where $S_1, \ldots, S_5$ are the sample variances of discriminant variables. The quantity '10' in $\mathrm{diag}(10S_1, \ldots, 10S_5)$ is selected according to normality of kernel data. Here note that among a few numbers, '10' is the one that leads to good normality in a series of Kolmogorov-Smirnov tests. The leading 24 principal components accounting for 99% of kernel data variability are used. Classification results for FLDA and KFLDA are reported in Table 2. (We ran FLDA and KFLDA using both SAS, proc discrim, and Matlab, classify, and obtained same classification results. The reason for trying it out in SAS is that an acquainted and user-friendly software is important for social science researchers to be willing to use it as a standard data analysis tool.) Results for SVM are not reported, because SVM fails to identify both the low and high ability groups by assigning almost all test points to the majority regular-ability group. We have also checked on the

Table 2: Comparison of classification results

		FLDA			KFLDA		
	classified	1	2	3	1	2	3
true	1	13	9	0	12	10	0
	2	40	163	106	16	222	71
	3	0	11	15	0	11	15

normality of the original data and the kernel data. The p-values of Kolmogorov-Smirnov tests for kernel data are in general significantly larger than those for original data. Normal probability plots (Figures 10-15) presented below also reveal that kernel data are better normally distributed.
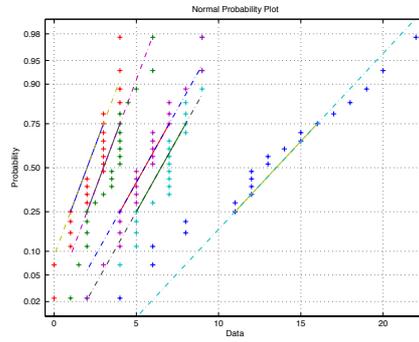
Figure 10: NP plots for $x_1, \ldots, x_5$-coordinates for group 1 data.
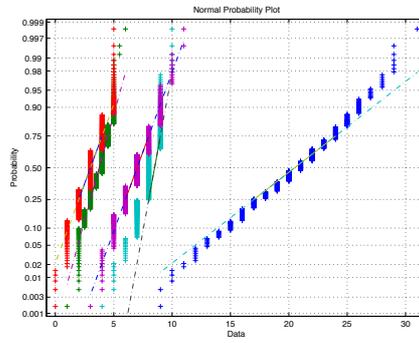

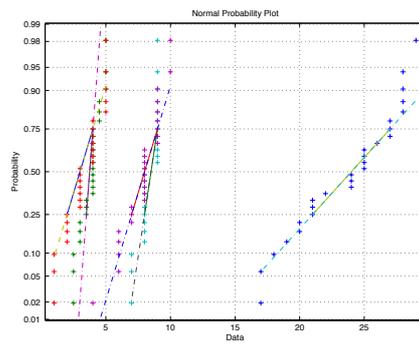Figure 11: NP plots for $x_1, \ldots, x_5$-coordinates for group 2 data.

Figure 12: NP plots for $x_1, \ldots, x_5$-coordinates for group 3 data.
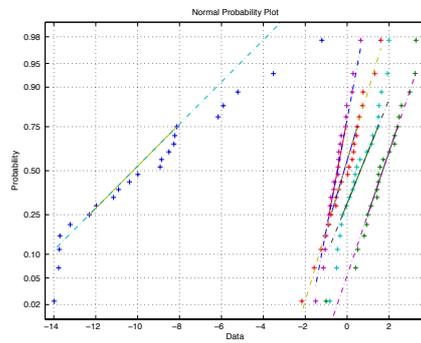
Figure 13: NP plots for the first 5 PCA-coordinates for group 1 kernel data.
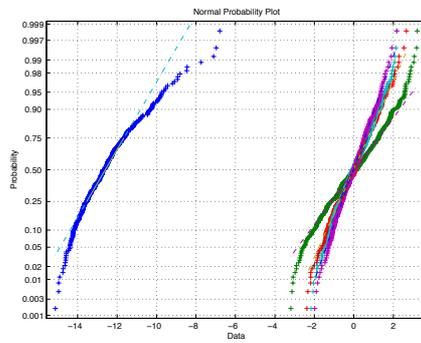

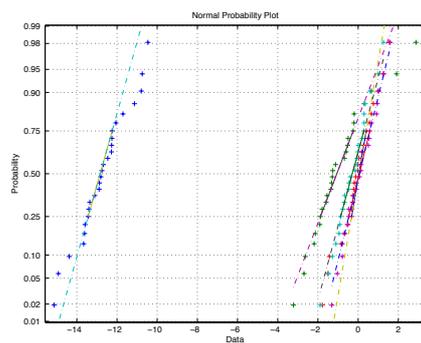Figure 14: NP plots for the first 5 PCA-coordinates for group 2 kernel data.

Figure 15: NP plots for the first 5 PCA-coordinates for group 3 kernel data.

16

Example 3 Consider a classification problem of four populations in R^2. The first variates have independent uniform distributions specified below

$$x_1^{(1)}, x_1^{(2)} \sim \text{uniform}(0, 3\pi), \quad x_1^{(3)}, x_1^{(4)} \sim \text{uniform}(3\pi, 6\pi).$$

The second variates are obtained by

$$x_2^{(i)} = \begin{cases} \sin(x_1^{(1)}) + 0.1\epsilon, & i = 1, \\ -\sin(x_1^{(2)}) + 0.1\epsilon, & i = 2, \\ \sin(x_1^{(3)}) + 0.1\epsilon, & i = 3, \\ -\sin(x_1^{(4)}) + 0.1\epsilon, & i = 4, \end{cases}$$

where ϵ is a standard normal noise. For each population a random sample of size 100 is drawn. These 400 sample points together with their population labels form the training set. Figure 16 depicts a typical data scatter.
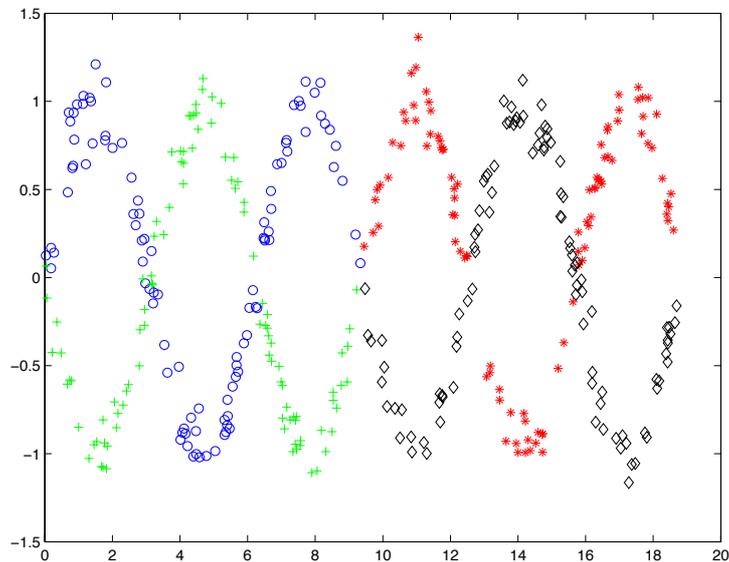
Figure 16: Data scatter plot for four populations.

We have tried out KFLDA and KFQDA (PCA-version) with Gaussian kernel and Epanechnikov kernel using leave-one-out cross validation on the training data. Various window widths are taken. Ten replicate runs are carried out. Average accuracy rates with standard errors in parentheses are reported in Tables 3 and 4. The number of leading principal components used and the average percentage (over 10 runs) of total variability accounted for by these leading coefficients are reported in the first column. From these results we see that the KFLDA is not as efficient as the KFQDA and sometimes can be inadequate. We also see that the KFQDA is more robust against the specification of window width and number of principal components. Choice of kernels has not much effect, though the Gaussian kernel performs slightly better.

17

Table 3: Comparison of KFLDA vs. KFQDA using Gaussian kernel.

no. PCs:	percentage	σ^2	KFLDA	KFQDA
8:	30.45%	0.25	0.6247 (0.0828)	0.8303 (0.0268)
10:	36.48%		0.6872 (0.0618)	0.8950 (0.0354)
15:	49.76%		0.8245 (0.0378)	0.9590 (0.0099)
25:	69.63%		0.9112 (0.0170)	—
35:	82.47%		0.9218 (0.0167)	—
120:	99.72%		0.9522 (0.0099)	—
150:	99.93%		0.9497 (0.0106)	—
200:	99.99%		0.9320 (0.0110)	—
8:	91.98%	5	0.5893 (0.0418)	0.9590 (0.0130)
10:	95.63%		0.5653 (0.0461)	0.9665 (0.0102)
15:	99.30%		0.9145 (0.0133)	0.9652 (0.0102)
25:	99.96%		0.9192 (0.0067)	0.9678 (0.0083)
35:	100%		0.9110 (0.0144)	0.9655 (0.0093)
100:	100%		0.9435 (0.0104)	—
120:	100%		***	—
8:	97.76%	10	0.5925 (0.0305)	0.9557 (0.0136)
10:	99.07%		0.5760 (0.0380)	0.9635 (0.0117)
15:	99.91%		0.8895 (0.0122)	0.9655 (0.0074)
25:	100%		0.9083 (0.0085)	0.9685 (0.0074)
35:	100%		0.9098 (0.0137)	—
75:	100%		0.9457 (0.0086)	—
80:	100%		***	—
8:	99.98%	50	0.6050 (0.0278)	0.9648 (0.0090)
10:	100%		0.5662 (0.0508)	0.9655 (0.0081)
15:	100%		0.8678 (0.0125)	0.9648 (0.0079)
25:	100%		0.9045 (0.0093)	0.9685 (0.0060)
35:	100%		0.9140 (0.0072)	—
40:	100%		0.9133 (0.0080)	—
45:	100%		***	—

— : indicates that at least one group sample covariance is singular.
*** : indicates that all four group sample covariances are singular.

Table 4: Comparison of KFLDA vs. KFQDA using Epanechnikov kernel.

no. PCs:	percentage	σ^2	KFLDA	KFQDA
10:	51.20%	1.5	0.7007 (0.0751)	0.9145 (0.0300)
15:	66.97%		0.8900 (0.0318)	0.9537 (0.0080)
20:	77.52%		0.9285 (0.0114)	0.9579 (0.0113)
100:	97.27%		0.9343 (0.0115)	—
150:	98.77%		0.9395 (0.0120)	—
200:	99.47%		0.9375 (0.0099)	—
10:	64.46%	2	0.6113 (0.0592)	0.9355 (0.0130)
15:	78.76%		0.8875 (0.0297)	0.9567 (0.0066)
25:	89.11%		0.9193 (0.0118)	0.9558 (0.0085)
100:	98.44%		0.9383 (0.0091)	—
150:	99.35%		0.9350 (0.0117)	—
200:	99.74%		0.9283 (0.0134)	—
10:	75.13%	2.5	0.5797 (0.0386)	0.9490 (0.0118)
15:	84.99%		0.9053 (0.0235)	0.9595 (0.0093)
25:	92.58%		0.9098 (0.0166)	0.9565 (0.0091)
100:	99.09%		0.9182 (0.0080)	—
150:	99.65%		0.9258 (0.0167)	—
200:	99.87%		0.9138 (0.0126)	—
10:	81.20%	3	0.5620 (0.0354)	0.9537 (0.0082)
15:	87.95%		0.8495 (0.0670)	0.9613 (0.0097)
25:	94.19%		0.9000 (0.0150)	0.9572 (0.0083)
100:	99.30%		0.9175 (0.0163)	—
150:	99.73%		0.9178 (0.0207)	—
200:	99.90%		0.9070 (0.0214)	—

— : indicates that at least one group sample covariance is singular.

Example 4 (Pen-based recognition of hand-written digits) This data set is taken from UCI machine learning data bases. The number of training instances is 7494 and the number of testing instances is 3498. There are 16 input measurements and a group label for each instance. Gaussian kernel with covariance $\text{diag}(10S_1, \ldots, 10S_{16})$ is used to make kernel data. The training data size is a bit larger than that in Examples 2 and 3, thus, instead of the PCA-based KFLDA, we have used the more economic random subset approach. Summary results of 10 runs are given below. As can be seen a random subset of a small

Table 5: Summary results for KFLDA-RS on Pen-digits data set

random subset size (proportion)	average accuracy (standard error)
100 (1.3%)	0.9574 (0.0042)
200 (2.6%)	0.9686 (0.0012)
300 (4.0%)	0.9725 (0.0011)
400 (5.3%)	0.9732 (0.0007)
500 (6.7%)	0.9739 (0.0013)
600 (8.0%)	0.9741 (0.0014)
1000 (13.3%)	0.9743 (0.0015)
1500 (20.0%)	0.9738 (0.0015)
2000 (26.7%)	0.9753 (0.0014)

portion is adequate.

Example 5 (Adult data set) This example serves to show that the KFDA training is applicable to a large data set with accuracy comparable to other classifiers. This data set is taken from UCI machine learning data bases. Instances with unknown values are removed, resulting in 45222 instances left. From the remaining instances we randomly select 2/3 (30148 instances) as training set and 1/3 (15074 instances) as test set. For each instance there are 14 input measurements, a mix of continuous and discrete measurements, and a group label determined by whether a person makes over 50K a year. Again, Gaussian kernel with covariance $\text{diag}(10S_1, \ldots, 10S_{14})$ is used to produce kernel data. We have used the more economic random subset approach. Prior probabilities are assigned according to training data group proportions, 0.75 for the group with annual earning less than or equal to 50K and 0.25 for the group with annual earning greater than 50K. Summary results of 10 runs are given in Table 6. Reported accuracy for various other algorithms ranging from

Table 6: Summary results for KFLDA-RS on Adult data set

random subset size	average accuracy (standard error)
301 (1%)	0.8512 (0.00095)
602 (2%)	0.8522 (0.00062)

78.58% to 85.95% can be found in Adult.names file in the UCI data bases. Accuracy rates for specific algorithms are: 84.46% for C4.5, 83.88% for Naive-Bayes, and 85.90% for NBTree. Our results are comparable to others, while the random subset approach significantly cuts down the computational load.

6   Theory of KFDA

The classical FDA can be derived as a maximum likelihood as well as a Bayes classifier for Gaussian populations in R^p. We will extend the classical theory to its kernel generalization. In Section 6.1 the KFDA is shown to be a maximum likelihood method under Gaussian assumption in the associated RKHS. In Section 6.2, under the same Gaussian assumption, the KFDA is shown to be a Bayes classifier. In Section 6.3 a theoretical justification of such a Gaussian assumption is provided. It is shown that, when data are represented via proper kernel functions in a RKHS, most low-dimensional projections of kernel data are approximately Gaussian under suitable conditions.

6.1   Maximum likelihood ratio of Gaussian measures

Let $(\mathcal{X}, \mathcal{B}, P_i)$, $i = 1, \ldots, k$, be probability measure spaces for populations $\pi_1, \ldots, \pi_k$. The data space $\mathcal{X}$ is embedded into an infinite dimensional space. Throughout Section 6 we assume that the underlying kernel $\kappa(x, u)$ is positive definite and trace class. Though the Gaussian and Epanechnikov kernels are not trace class on $R^p \times R^p$ (since $\int_{R^p} \kappa(x, x)dx = \infty$), it does not hinder us from using them. The reason is that as long as the data space $\mathcal{X}$ is compact, then $\int_{\mathcal{X}} \kappa(x, x)dx < \infty$; i.e., $\kappa(x, u)$ is trace class if restricted to $\mathcal{X} \times \mathcal{X}$. Recall the transformation $\Gamma : x \to \kappa(x, \cdot)$, which embeds the original data space into a new one equipped with a richer topological structure. Let $X^{(i)}$ denote the random variable for $(\mathcal{X}, \mathcal{B}, P_i)$, $i = 1, \ldots, k$. Let

$$m_i(\cdot) = E\kappa(X^{(i)}, \cdot) \text{ and } \Lambda_i(x, u) = cov\{\kappa(X^{(i)}, x), \kappa(X^{(i)}, u)\}. \tag{14}$$

The covariance kernel Λ_i can induce an operator on $\mathcal{H}_\kappa$ as defined below:

$$\Lambda_i f(x) = \langle \Lambda_i(x, \cdot), f(\cdot) \rangle_{\mathcal{H}_\kappa} \ \ \forall f \in \mathcal{H}_\kappa.$$

The induced operator is bounded, linear, nonnegative-definite, self-adjoint and trace-class. Such an operator is called a covariance operator or a dispersion operator (see, e.g., Vakhania *et al.*, 1987, and Rao and Varadarajan, 1963).

Definition 3 (Covariance operator) *A covariance operator in $\mathcal{H}_\kappa$ is defined to be an operator, which is bounded, linear, nonnegative-definite, self-adjoint and trace class.*

We will not distinguish between a covariance kernel and its induced covariance operator, unless it is necessary. The mean functions m_i's and the covariance operators Λ_i's satisfy the following properties (see Vakhania *et al.*, 1987):

$$E\langle h_i, f \rangle_{\mathcal{H}_\kappa} = \langle m_i, f \rangle_{\mathcal{H}_\kappa} = \langle Eh_i, f \rangle_{\mathcal{H}_\kappa}, \forall f \in \mathcal{H}_\kappa, \text{ and} \tag{15}$$

21

$$cov\{\langle h_i, f \rangle_{\mathcal{H}_\kappa}, \langle h_i, g \rangle_{\mathcal{H}_\kappa}\} = \langle \Lambda_i f, g \rangle_{\mathcal{H}_\kappa} = \langle f, \Lambda_i g \rangle_{\mathcal{H}_\kappa}, \forall f, g \in \mathcal{H}_\kappa, \qquad (16)$$

where $h_i = \kappa(X^{(i)}, \cdot)$ denotes a random element in $\mathcal{H}_\kappa$ induced by Γ from the probability measure P_i. Identity (15) says that we may exchange the order of expectation and $\mathcal{H}_\kappa$-inner product, and identity (16) says that we may exchange the order of covariance operator and $\mathcal{H}_\kappa$-inner product.

Below we give definitions of linear and quadratic classifiers in $\mathcal{H}_\kappa$.

Definition 4 (linear classifier) *Consider a binary classification in $\mathcal{H}_\kappa$. We say that a classifier is linear if and only if its decision boundary is given by*

$$\ell(h) + b = 0,$$

where $\ell(\cdot)$ is a bounded linear functional, b is a real scalar and h is a sample element in $\mathcal{H}_\kappa$.

By Riesz Representation Theorem, for each linear functional $\ell(\cdot)$ there exists a unique $g \in \mathcal{H}_\kappa$ such that the decision boundary is given by

$$\langle g, h \rangle_{\mathcal{H}_\kappa} + b = 0. \qquad (17)$$

The idea of KFLDA is to look for an optimal separating hyperplane which has $g \in \mathcal{H}_\kappa$ as its *functional normal direction*. The optimality is in the following sense. When data patterns (conveyed in realizations $\Gamma(x_j)$, $j = 1 \ldots, n$) are projected along g, the group centers are far apart, while the spread within each group is small causing the overlap of these two groups to be as small as possible in this functional direction.

Definition 5 (quadratic classifier) *Consider a binary classification in $\mathcal{H}_\kappa$. We say that a classifier is quadratic if and only if its decision boundary is given by*

$$\langle h, Ah \rangle_{\mathcal{H}_\kappa} + \ell(h) + b = 0,$$

where A is a symmetric (in the sense that $\langle h, Ah \rangle_{\mathcal{H}_\kappa} = \langle Ah, h \rangle_{\mathcal{H}_\kappa}$) linear operator, $\ell(\cdot)$ is a bounded linear functional, b is a real scalar and h is a sample element in $\mathcal{H}_\kappa$.

The idea of KFQDA is to look for an optimal quadratic separating hypersurface. The optimality is in the following sense. When data patterns are processed through the quadratic form, $\langle \Gamma(x_j), A\Gamma(x_j) \rangle_{\mathcal{H}_\kappa} + \ell(\Gamma(x_j))$, the group centers are far apart, while the spread within each group is small causing the overlap of these two groups to be as small as possible.

Later in Section 6.3 we will show that, if the embedding sample space is properly chosen, the underlying populations' distributions in this new data space can be well approximated by Gaussian measures under some mild conditions. Both the KFLDA and KFQDA can be obtained from likelihood ratio of Gaussian measures on RKHS. Since it is a likelihood ratio based criterion, the k-group KFDA can be broken down into $k - 1$ many binary classifications. Thus, it is sufficient to investigate the likelihood ratio of two Gaussian measures. The linear case is due to Grenander (1950) and the quadratic case is due to Rao and Varadarajan (1963).

Theorem 1 (Grenander, 1950; Rao & Varadarajan, 1963) *Assume $P_{1,\mathcal{H}_\kappa}$ and $P_{2,\mathcal{H}_\kappa}$ are two equivalent Gaussian measures on $\mathcal{H}_\kappa$ with mean functions $m_1(x)$ and $m_2(x)$ and nonsingular covariance operators Λ_1 and Λ_2. Let $L_{2,1} = \log(dP_{2,\mathcal{H}_\kappa}/dP_{1,\mathcal{H}_\kappa})$ and h be an element in $\mathcal{H}_\kappa$. Let $m_a = (m_1 + m_2)/2$ and $m_d = m_2 - m_1$.*

(1) For the case $\Lambda_1 = \Lambda_2 = \Lambda$, a necessary and sufficient condition that the log-likelihood ratio $L_{2,1}$ be linear is that $m_d \in R(\Lambda)$, where $R(\Lambda)$ is the range of Λ. The log-likelihood ratio is then given by

$$L_{2,1}(h) = \langle h, \Lambda^{-1} m_d \rangle_{\mathcal{H}_\kappa} - \langle m_a, \Lambda^{-1} m_d \rangle_{\mathcal{H}_\kappa}. \tag{18}$$

(2) For the case $\Lambda_1 \neq \Lambda_2$, suppose that (a) $R(\Lambda_1) = R(\Lambda_2)$, (b) M^*M defines a bounded and trace class linear operator on $\mathcal{H}_\kappa$, where $M = (\Lambda_2^{-1} - \Lambda_1^{-1})\Lambda_1^{1/2}$ and (c) $m_d \in R(\Lambda_1)$. Then the closure A of $\Lambda_2^{-1} - \Lambda_1^{-1}$ exists and is a closed densely defined symmetric operator such that $P_1\{D(A) + m_a\} = 1$, where $D(A)$ is the domain of A. The log-likelihood ratio is then given by

$$\begin{aligned}
L_{2,1}(h) &= -\frac{1}{2}\langle h - m_a, A(h - m_a)\rangle_{\mathcal{H}_\kappa} + \frac{1}{2}\langle h - m_a, (\Lambda_2^{-1} + \Lambda_1^{-1})m_d\rangle_{\mathcal{H}_\kappa} \\
&\quad -\frac{1}{8}\langle m_d, Am_d\rangle_{\mathcal{H}_\kappa} - \frac{1}{2}\log(|S|),
\end{aligned}$$

where S is the operator satisfying the equation $\Lambda_2 = \Lambda_1^{1/2} S \Lambda_1^{1/2}$.

By abusing the notation we re-write the above log-likelihood ratio in a conventional form resembling the Mahalanobis distance formulation (up to a constant term):

$$\begin{aligned}
L_{2,1}(h) &= \frac{1}{2}\left\{\langle h - m_1, \Lambda_1^{-1}(h - m_1)\rangle_{\mathcal{H}_\kappa} - \langle h - m_2, \Lambda_2^{-1}(h - m_2)\rangle_{\mathcal{H}_\kappa}\right\} \\
&\quad + \frac{1}{2}\log(|\Lambda_1|/|\Lambda_2|). \tag{19}
\end{aligned}$$

Note that in the infinite dimensional case Λ_1^{-1} and Λ_2^{-1} are unbounded operators and are not defined for a large class of elements in $\mathcal{H}_\kappa$ and that the determinants of Λ_1 and Λ_2 are zero. Expression (19), though valid for the case of a finite dimensional Hilbert space, is not a rigorous one for infinite dimensional case, but is merely made out of convenience for an explanatory purpose. To separate two Gaussian populations in $\mathcal{H}_\kappa$, the log-likelihood ratio in Theorem 1 leads to an ideal optimal linear decision boundary for case (1), and an ideal optimal quadratic decision boundary for case (2). There are parameters including mean function(s) and covariance operator(s) involved in the log-likelihood ratio, which have to be estimated from the data. Below we derive their maximum likelihood estimates.

Theorem 2 (Maximum likelihood estimates) *Assume that $\{X_j\}_{j=1}^n$ are iid observations from a probability measure $(\mathcal{X}, \mathcal{B}, P)$. Assume the Γ-induced probability measure $P_{\mathcal{H}_\kappa}$ is Gaussian with mean function $m(\cdot)$ and nonsingular covariance operator Λ. Then, for arbitrary functions g and f in $\mathcal{H}_\kappa$, the maximum likelihood estimate for $\langle g, m \rangle_{\mathcal{H}_\kappa}$ is given by $\langle g, \hat{m} \rangle_{\mathcal{H}_\kappa}$ with*

$$\hat{m}(x) = \frac{1}{n}\sum_{j=1}^n \kappa(X_j, x), \tag{20}$$

and the maximum likelihood estimate for $\langle g, \Lambda f \rangle_{\mathcal{H}_\kappa}$ is given by $\langle g, \hat{\Lambda} f \rangle_{\mathcal{H}_\kappa}$ with

$$\hat{\Lambda}(x, u) = \frac{1}{n} \sum_{j=1}^{n} (\kappa(X_j, x) - \hat{m}(x))(\kappa(X_j, u) - \hat{m}(u)). \tag{21}$$

In particular, for given x and u, by taking g and f the evaluation functionals at x and u respectively, the MLEs for $m(x)$ and $\Lambda(x, u)$ are given by (20) and (21), respectively.

For multiple populations sharing a common covariance operator, we pool together sample covariance estimates from all populations according to their sizes to get a pooled single estimate. Theorems 1 and 2 lead to the maximum likelihood method that coincides with either the KFLDA or the KFQDA depending on if a common covariance operator is assumed.

6.2 Bayesian discrimination of Gaussian measures

In constructing a classifier, it is desired to minimize the probability of misclassification. Let $R_i \subset \mathcal{H}_\kappa$ be disjoint regions of classification as from group i satisfying $\mathcal{H}_\kappa = \cup_{i=1}^{k} R_i$ and let $R_i^c = \mathcal{H}_\kappa \setminus R_i$. Let C_i denote the cost of misclassifying an object from group i into other groups. Let q_i denote the prior probability of an observation coming from group i. The expected loss is then given by

$$\sum_{i=1}^{k} q_i C_i P_{i,\mathcal{H}_\kappa}(R_i^c). \tag{22}$$

Definition 6 (Bayes classifier) *A classifier, consisting of disjoint classification regions* $\{R_i\}_{i=1}^{k}$ *and satisfying* $\mathcal{H}_\kappa = \cup_{i=1}^{k} R_i$, *that minimizes (22) for prior probabilities* $q = (q_1, \ldots, q_k)$ *and costs* $C = (C_1, \ldots, C_k)$ *is called a Bayes classifier in* $\mathcal{H}_\kappa$.

Theorem 3 (Bayes classifier) *Assume* $P_{i,\mathcal{H}_\kappa}$, $i = 1, \ldots, k$, *are equivalent probability measures on* $\mathcal{H}_\kappa$. *Then, the classifier consisting of disjoint classification regions*

$$R_i = \{h : \log(q_i C_i) + L_{i,1}(h) \geq \log(q_j C_j) + L_{j,1}(h), \ j = 1, \ldots, k\}$$

is Bayes, where $L_{i,1} = \log(dP_{i,\mathcal{H}_\kappa}/dP_{1,\mathcal{H}_\kappa})$. *(For groups with ties, it does not matter which group h is assigned to.)*

Corollary 4 *If* $P_{i,\mathcal{H}_\kappa}$, $i = 1, \ldots, k$, *are equivalent Gaussian measures satisfying assumptions in Theorem 1, the Bayes classifier under* $C_1 = \cdots = C_k$ *is a likelihood ratio method. Furthermore, with parameters estimated by maximum likelihood the Bayes classifier coincides with the KFDA. When C_i are not all equal, we simply incorporate them into q_i, then the Bayes and the KFDA still coincide.*

6.3 Justification of Gaussian assumption

Results in the previous two sections are based on the Gaussian assumption on a RKHS. In this section we provide a justification for it.

6.3.1 Single population: the non-centered case

Let $\Gamma(x_1), \ldots, \Gamma(x_n)$ be (nonrandom) functions in $\mathcal{H}_\kappa$. They form the data set. The kernel window width σ depends on n and so do the kernel κ and the associated Hilbert space $\mathcal{H}_\kappa$. Notations σ_n, κ_n and $\mathcal{H}_n$ will be used from now on to indicate their dependence on n. Let κ_o denote the baseline kernel with window width one. Then $\kappa_n(x, u) = \kappa_o(x/\sigma_n, u/\sigma_n)/\sigma_n^p$. The size of σ_n controls the resolution of the associated Hilbert space. The smaller σ_n is, the finer resolution the space $\mathcal{H}_n$ has. The resolution notion can be easily seen in the splines and wavelets associated Hilbert spaces via a sequence of nested Hilbert spaces. The window width σ_n should be decreasing to zero, as the sample size approaches infinity. Suppose that there exists a constant $\tau^2 > 0$ such that for any $\epsilon > 0$, as $n \to \infty$,

$$\frac{1}{n} \, \mathrm{card}\{1 \le j \le n : |\sigma_n^p \|\Gamma(x_j)\|_{\mathcal{H}_n}^2 - \tau^2| > \epsilon\} \to 0,$$

$$\frac{1}{n^2} \, \mathrm{card}\{1 \le j, j' \le n : \sigma_n^p |\langle \Gamma(x_j), \Gamma(x_{j'})\rangle_{\mathcal{H}_n}| > \epsilon\} \to 0,$$

where $\sigma_n \to 0$. By the reproducing property of kernel, the two conditions above are equivalent to

$$\frac{1}{n} \, \mathrm{card}\{1 \le j \le n : |\sigma_n^p \kappa_n(x_j, x_j) - \tau^2| > \epsilon\} \to 0, \tag{23}$$

$$\frac{1}{n^2} \, \mathrm{card}\{1 \le j, j' \le n : \sigma_n^p |\kappa_n(x_j, x_{j'})| > \epsilon\} \to 0. \tag{24}$$

Condition (23) says that most kernel data $\{\sqrt{\sigma_n^p}\Gamma(x_j)\}_{j=1}^n$ have $\mathcal{H}_n$-norm near τ^2. Condition (24) says that most kernel data are nearly orthogonal in $\mathcal{H}_n$. Let h be a random element from a Gaussian measure with zero mean function and the covariance operator κ_n. Kernel data $\{\sqrt{\sigma_n^p}\Gamma(x_j)\}_{j=1}^n$ projected along the random direction h have values given by

$$\sqrt{\sigma_n^p}\langle h, \Gamma(x_1)\rangle_{\mathcal{H}_n}, \cdots, \sqrt{\sigma_n^p}\langle h, \Gamma(x_n)\rangle_{\mathcal{H}_n}.$$

Let $\theta_n(h)$ be the empirical distribution of this sequence, assigning probability mass n^{-1} to each $\sqrt{\sigma_n^p}\langle h, \Gamma(x_j)\rangle_{\mathcal{H}_n}$.

Theorem 5 *Under conditions (23) and (24), as $n \to \infty$, the empirical distribution $\theta_n(h)$ converges weakly to $N(0, \tau^2)$ in probability.*

Theorem 5 is established for a one-dimensional projection along any random direction h and it can be extended to an m-dimensional projection along random directions $h_1, \ldots, h_m$ for an arbitrary but fixed m. The KFDA's effective working subspace is actually a low-dimensional one. Theorem 5 says that most low-dimensional projections of kernel data can be well approximated by a Gaussian distribution.

The following conditions imply (23) and (24):

$$\frac{\sigma_n^p}{n} \sum_{j=1}^n \kappa_n(x_j, x_j) \to \tau^2, \tag{25}$$

$$\frac{1}{n} \text{ card } \{1 \le j \le n : |\sigma_n^p \kappa_n(x_j, x_j) - \tau^2| > \epsilon\} \to 0, \tag{26}$$

$$\frac{\sigma_n^{2p}}{n^2} \sum_{j,j'=1}^{n} \kappa_n^2(x_j, x_{j'}) \to 0. \tag{27}$$

Theorem 6 *Assume conditions (25)-(27) hold. Let $t_n^2(h) = n^{-1} \sum_{j=1}^{n} \sigma_n^p \langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}^2$ and let $\theta_n^1(h)$ be the scaled empirical distribution, assigning probability mass n^{-1} to each of $\sqrt{\sigma_n^p} \langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n} / t_n$. Then*
(1) the empirical second moment $t_n^2(h)$ converges to τ^2 in probability, and
(2) the scaled empirical distribution $\theta_n^1(h)$ converges weakly to $N(0,1)$ in probability.

6.3.2 Single population: the centered case

Next we consider the asymptotic distribution of centered data projected along a random direction h from $N(0, \kappa_n)$. Let

$$\bar{\Gamma} = n^{-1} \sum_{j=1}^{n} \Gamma(x_j), \ \ \tilde{\Gamma}(x_j) = \Gamma(x_j) - \bar{\Gamma}, \ \ a_n(h) = \sqrt{\sigma_n^p} \langle h, \bar{\Gamma} \rangle_{\mathcal{H}_\kappa} \ \text{ and}$$

$$s_n^2(h) = n^{-1} \sum_{j=1}^{n} [\sqrt{\sigma_n^p} \langle h, \Gamma(x_j) \rangle_{\mathcal{H}_\kappa} - a_n]^2 = n^{-1} \sigma_n^p \sum_{j=1}^{n} \langle h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_\kappa}^2.$$

Let $\theta_n^0(h)$ be the centered empirical distribution, assigning probability mass n^{-1} to each $\sqrt{\sigma_n^p} \langle h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_\kappa}$. Define

$$\kappa_{j\cdot} = n^{-1} \sum_{j'=1}^{n} \kappa_n(x_j, x_{j'}), \ \ \kappa_{\cdot j'} = n^{-1} \sum_{j=1}^{n} \kappa_n(x_j, x_{j'}), \ \ \kappa_{\cdot\cdot} = n^{-2} \sum_{j,j'=1}^{n} \kappa_n(x_j, x_{j'}).$$

Conditions below are centered versions of conditions (25)-(27):

$$\frac{\sigma_n^p}{n} \sum_{j=1}^{n} \|\tilde{\Gamma}(x_j)\|_{\mathcal{H}_n}^2 \to \tau^2,$$

$$\frac{1}{n} \text{ card } \{1 \le j \le n : |\sigma_n^p \|\tilde{\Gamma}(x_j)\|_{\mathcal{H}_n}^2 - \tau^2| > \epsilon\} \to 0,$$

$$\frac{\sigma_n^{2p}}{n^2} \sum_{j,j'=1}^{n} \langle \tilde{\Gamma}(x_j), \tilde{\Gamma}(x_{j'}) \rangle_{\mathcal{H}_n}^2 \to 0.$$

By the kernel reproducing property, conditions above are equivalent to conditions (28)-(30).

$$\frac{\sigma_n^p}{n} \sum_{j=1}^{n} [\kappa_n(x_j, x_j) - \kappa_{\cdot\cdot}] \to \tau^2, \tag{28}$$

$$\frac{1}{n} \text{ card } \{1 \le j \le n : |\sigma_n^p (\kappa_n(x_j, x_j) - 2\kappa_{j\cdot} + \kappa_{\cdot\cdot}) - \tau^2| > \epsilon\} \to 0, \tag{29}$$

$$\frac{\sigma_n^{2p}}{n^2} \sum_{j,j=1}^{n} [\kappa_n(x_j, x_{j'}) - \kappa_{j\cdot} - \kappa_{\cdot j'} + \kappa_{\cdot\cdot}]^2 \to 0. \tag{30}$$

Theorem 7 *Assume conditions (28)-(30) hold and let h be a random element from $N(0, \kappa_n)$. Then*
(1) the empirical variance $s_n^2(h)$ converges to τ^2 in probability, and
(2) the centered empirical distributions $\theta_n^0(h)$ converges weakly to $N(0, \tau^2)$ in probability.

Remark 5 *In either the centered or non-centered case, the asymptotic empirical distribution does not depend on the projection direction h. This phenomenon indicates that the kernel transformed data $\Gamma(x_j)$ distribution looks spherically symmetrically over $\mathcal{H}_n$ when n is large. Here the spherical symmetry is with respect to the $\mathcal{H}_n$ topology rather than the L_2 topology.*

6.3.3 Multiple populations

Suppose training inputs $x_1, \ldots, x_n$ are independent samples from one of $\pi_1, \ldots, \pi_k$ with corresponding group label y_j. With kernel transformation, the new data set consists of $\{(\Gamma(x_j), y_j)\}_{j=1}^n$. Let $\bar{\Gamma}_i(u) = n_i^{-1} \sum_{j \in I_i} \kappa_n(x_j, u)$ and

$$\tilde{\Gamma}(x_j)(u) = \kappa_n(x_j, u) - \bar{\Gamma}_i(u), \text{ if corresponding group label } y_j = i.$$

When conditions (28)-(30) are met, Theorem 7 is valid and can be used to classify data into different populations. By projecting the centered data from all groups along a common random direction $h \sim N(0, \kappa_n)$, Theorem 7 says that the empirical distribution for the ith group, by assigning probability mass n_i^{-1} to each $\sqrt{\sigma_n^p} \langle h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_n}$ for $j \in I_i$, converges weakly to a normal distribution. We would like to see mean functions and/or the variances are population dependent. Note that Theorem 7 is established under conditions (28)-(30). However, if conditions (23)-(24) are also met, then by Theorem 5 all these k empirical distributions $\theta_{n_i}(h)$ converge weakly to an identical distribution $N(0, \tau^2)$ and so do their centered empirical distributions. Thus these k populations are not distinguishable by asymptotic empirical distributions based on kernel data. Thus, ideally we would like to see that conditions (28)-(30) are met but condition (24) is violated. If so, then these k populations can be discriminated by the method of KFDA. Otherwise, they are not distinguishable by such an approach. Next we will check that under what circumstance these conditions are met or violated.

Proposition 1 *Let $X, X_1, X_2, X_3, \ldots$ be an iid sequence from a continuous distribution having probability density function $p(x)$. Assume that $\kappa_o(s, t)$ is decreasing to zero as $\|s - t\| \to \infty$. Also assume that $\sigma_n \to 0$. Let $r = \lim_{n \to \infty} n \sigma_n^p$.*
(1) If $0 \leq r < \infty$, then for any $\epsilon > 0$ and any fixed positive integer i we have

$$\lim_{n \to \infty} P\{ \text{ card}\{1 \leq j \leq n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} = i \} = \frac{e^{-rq}(rq)^i}{i!},$$

where $q(\epsilon) = \lim_{n \to \infty} \sigma_n^{-p} P\{|\kappa_o(X_j/\sigma_n, X/\sigma_n)| > \epsilon\} > 0$. In particular, if $r = 0$, we have $\lim_{n \to \infty} \text{card}\{1 \leq j \leq n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} = 0$ in probability.
(2) If $r = \infty$, then for any $\epsilon > 0$ we have $\text{card}\{1 \leq j \leq n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} \to \infty$ in probability.

Result (1) implies that if $\sigma_n \to 0$ at a rate fast enough so that $r < \infty$, then

$$\frac{1}{n} \operatorname{card}\{1 \le j \le n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} \to 0 \text{ in probability},$$

which implies

$$\frac{1}{n^2} \operatorname{card}\{1 \le j, j' \le n : \sigma_n^p |\kappa_n(X_j, X_{j'})| > \epsilon\} \to 0 \text{ in probability}.$$

However, if $\sigma_n \to 0$ at a rate not so fast that $r = \infty$, then condition (24) is violated in the sense of result (2) of Proposition 1. Though result (2) of Proposition 1 does not quarantee the almost sure failure of condition (24), it says that for each j'th column $\operatorname{card}\{1 \le j \le n : \sigma_n^p |\kappa_n(X_j, X_{j'})| > \epsilon\} \to \infty$ in probability.

Theorem 8 *Let $X, X_1, X_2, X_3, \ldots$ be an iid sequence from a continuous distribution having probability density function $p(x)$. Assume that $\kappa_o(s, t)$ is decreasing to zero as $\|s - t\| \to \infty$ and that $\sigma_n \to 0$ as $n \to \infty$. Then*
(1) conditions (25) and (26) hold almost surely and so do conditions (28) and (29) with $\tau^2 = \lim_{n \to \infty} E\kappa_o(X/\sigma_n, X/\sigma_n)$;
(2) condition (30) also holds almost surely.

Remark 6 *Polynomial kernels do not satisfy the tail decay property and hence Proposition 1 and Theorem 8 do not apply to polynomial kernels.*

Remark 7 *From Proposition 1 and Theorem 8, we see that the kernel employed should have tail decay and an ideal window width should be controlled in a way that $n\sigma_n^p \to \infty$ and $\sigma_n \to 0$, so that, for iid sequence $X_1, X_2, X_3, \ldots$ from a continuous distribution, conditions (25)-(26) and conditions (28)-(30) hold almost surely, while condition (24) is violated.*

Remark 8 *Proposition 1 and Theorem 8 are established for continuous distribution. However, some of the data sets in Section 5 are mixtures of continuous and discrete types. From our empirical experience low-dimensional projections of kernel data can still be well approximated by Gaussian distribution for mixed-type input variables, as long as the projections are not onto a purely discrete subspace.*

Acknowledgment

Appendix

Proof of Theorem 2: For an arbitrary pair $f, g \in \mathcal{H}_\kappa$, the random vector given by $(\langle g, \kappa(X, \cdot) \rangle_{\mathcal{H}_\kappa}, \langle f, \kappa(X, \cdot) \rangle_{\mathcal{H}_\kappa})$ has a 2-dimensional Gaussian distribution by Definition 2. From (15), the mean for the first variate is $E\langle g, \kappa(X, \cdot) \rangle_{\mathcal{H}_\kappa} = \langle g, m \rangle_{\mathcal{H}_\kappa}$; and from (16), the covariance is $cov\{\langle g, \kappa(X, \cdot) \rangle_{\mathcal{H}_\kappa}, \langle f, \kappa(X, \cdot) \rangle_{\mathcal{H}_\kappa}\} = \langle g, \Lambda f \rangle_{\mathcal{H}_\kappa}$. The MLEs for a 2-dimensional Gaussian are clear. □

Proof of Theorem 3: We are aiming to find disjoint regions R_i so as to minimize (22). Assume for a moment that $C_1 = \cdots = C_k$. For a given h we minimize the expected loss by assigning h to the group having the highest conditional probability, i.e.,

$$R_i = \{h : q_i dP_{i,\mathcal{H}_\kappa}(h) \geq q_j dP_{j,\mathcal{H}_\kappa}(h) \text{ for } 1 \leq j \leq k \text{ and } j \neq i\}.$$

Or equivalently, $R_i = \{h : \log(q_i) + L_{i,1}(h) = \max_{j=1}^k \log(q_j) + L_{j,1}(h)\}$. Taking cost into account, the region of classification is then given by $R_i = \{h : \log(q_i C_i) + L_{i,1}(h) = \max_{j=1}^k \log(q_j C_j) + L_{j,1}(h)\}$. □

Proof of Theorem 5: In this proof and beyond, var_h and E_h are respectively the variance and expectation with respect to the distribution of the random element h. The characteristic function of $\theta_n(h)$ is

$$\phi_n(h, t) = n^{-1} \sum_{j=1}^n \exp\{it\sqrt{\sigma_n^p}\langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}\}.$$

Note that $var_h\{\langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}\} = \langle \Gamma(x_j), \kappa_n \Gamma(x_j) \rangle_{\mathcal{H}_n} = \kappa_n(x_j, x_j)$, where the first equality holds by property (16) and the second equality holds by the reproducing property of κ_n. Then, by condition (23),

$$E_h\phi_n(h, t) = n^{-1} \sum_{j=1}^n \exp\{-t^2\sigma_n^p \kappa_n(x_j, x_j)/2\} \rightarrow \exp\{-t^2\tau^2/2\}.$$

Likewise, since

$$var_h\{\langle h, \Gamma(x_j) - \Gamma(x_{j'}) \rangle_{\mathcal{H}_n}\} = \kappa_n(x_j, x_j) + \kappa_n(x_{j'}, x_{j'}) - 2\kappa_n(x_j, x_{j'}),$$

then

$$E_h\{|\phi_n(h, t)|^2\} = n^{-2} \sum_{j,j'=1}^n E_h\{\exp(it\sqrt{\sigma_n^p}\langle h, \Gamma(x_j) - \Gamma(x_{j'}) \rangle_{\mathcal{H}_n})\}$$

$$= n^{-2} \sum_{j,j'=1}^n \exp\{-t^2\sigma_n^p[\kappa_n(x_j, x_j) + \kappa_n(x_{j'}, x_{j'}) - 2\kappa_n(x_j, x_{j'})]/2\} \rightarrow \exp\{-t^2\tau^2\}$$

by conditions (23) and (24). Thus, by Chebychev's inequality, $\phi_n(h, t) \rightarrow \exp\{-t^2\tau^2/2\}$ in probability. Using Lemma 2.2 of Diaconis and Freedman (1984), we may conclude that $\theta_n(h)$ converges weakly to $N(0, \tau^2)$ in probability. □

Proof of Theorem 6: (1) By condition (25)

$$E_h t_n^2 = n^{-1} \sum_{j=1}^n E_h \sigma_n^p \langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}^2 = n^{-1}\sigma_n^p \sum_{j=1}^n \kappa_n(x_j, x_j) \rightarrow \tau^2.$$

Since $(\langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}, \langle h, \Gamma(x_{j'}) \rangle_{\mathcal{H}_n})$ is a bivariate normal with zero mean and covariance matrix:

$$\left[\begin{array}{cc} \kappa_n(x_j, x_j) & \kappa_n(x_j, x_{j'}) \\ \kappa_n(x_j, x_{j'}) & \kappa_n(x_j, x_j) \end{array} \right],$$

similar to Lemma 2.3 in Diaconis and Freedman (1984), we have

$$E_h[\langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}^2 \langle h, \Gamma(x_{j'}) \rangle_{\mathcal{H}_n}^2] = 2[\kappa_n(x_j, x_{j'})]^2 + \kappa_n(x_j, x_j)\kappa_n(x_{j'}, x_{j'}).$$

By conditions (26) and (27)

$$
\begin{aligned}
E_h t_n^4 &= n^{-2}\sigma_n^{2p} E_h[\sum_{j=1}^{n} \langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}^2]^2 = n^{-2}\sigma_n^{2p} \sum_{j,j'=1}^{n} E_h[\langle h, \Gamma(x_j) \rangle_{\mathcal{H}_n}^2 \langle h, \Gamma(x_{j'}) \rangle_{\mathcal{H}_n}^2] \\
&= 2n^{-2}\sigma_n^{2p} \sum_{j,j'=1}^{n} [\kappa_n(x_j, x_{j'})]^2 + n^{-2}\sigma_n^{2p} \sum_{j,j'=1}^{n} \kappa_n(x_j, x_j)\kappa_n(x_{j'}, x_{j'}) \to \tau^4.
\end{aligned}
$$

Thus, $t_n^2 \to \tau^2$ in probability.

(2) By Theorem 5 and Theorem 6-(1), $\theta_n^1 \to N(0,1)$ weakly in probability can be obtained by Slutzky's lemma. $\qquad \square$

Proof of Theorem 7: (1) By condition (28)

$$E_h s_n^2 = n^{-1} \sum_{j=1}^{n} E_h \langle \sqrt{\sigma_n^p} h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_\kappa}^2 = n^{-1}\sigma_n^p \sum_{j=1}^{n} [\kappa(x_j, x_j) - \kappa_{..}] \to \tau^2.$$

Similarly, by conditions (26) and (27), we have $E_h s_n^4 \to \tau^4$. Thus, $s_n^2 \to \tau^2$ in probability.

(2) The characteristic function of $\theta_n^0(h)$ is

$$\phi_n(h, t) = n^{-1} \sum_{j=1}^{n} \exp\{it\langle \sqrt{\sigma_n^p} h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_\kappa}\}.$$

Since $var_h\{\langle h, \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_n}\} = \langle \tilde{\Gamma}(x_j), \kappa_n \tilde{\Gamma}(x_j) \rangle_{\mathcal{H}_n} = \kappa_n(x_j, x_j) - 2\kappa_{j.} + \kappa_{..}$, then

$$E_h \phi_n(h, t) = n^{-1} \sum_{j=1}^{n} \exp\{-t^2 \sigma_n^p[\kappa(x_j, x_j) - 2\kappa_{j.} + \kappa_{..}]/2\} \to \exp\{-t^2\tau^2/2\}$$

by condition (28). Likewise, since $var_h\{\langle h, \tilde{\Gamma}(x_j) - \tilde{\Gamma}(x_{j'}) \rangle_{\mathcal{H}_\kappa}\} = \|\tilde{\Gamma}(x_j) - \tilde{\Gamma}(x_{j'})\|_{\mathcal{H}_\kappa}^2$, then

$$
\begin{aligned}
& E_h\{|\phi_n(h, t)|^2\} \\
&= n^{-2} \sum_{j,j'=1}^{n} E_h\{\exp(it\langle \sqrt{\sigma_n^p} h, \tilde{\Gamma}(x_j) - \tilde{\Gamma}(x_{j'}) \rangle_{\mathcal{H}_\kappa})\} \\
&= n^{-2} \sum_{j,j'=1}^{n} \exp\{-t^2 \sigma_n^p \|\tilde{\Gamma}(x_j) - \tilde{\Gamma}(x_{j'})\|_{\mathcal{H}_\kappa}^2/2\} \to \exp\{-t^2\tau^2\}
\end{aligned}
$$

by conditions (28)-(30). Thus, by Chebychev's inequality, $\phi_n(h, t) \to \exp\{-t^2\tau^2/2\}$ in probability. We may conclude that $\theta_n^0(h)$ converges weakly to $N(0, \tau^2)$ in probability. $\qquad \square$

Lemma 1 (Poisson approximation) *Suppose for each n, $Z_{n1}, \ldots, Z_{nr_n}$ are independent random variables, where each Z_{nk} is a Bernoulli trial with probability of success p_{nk}. If $\lim_{n\to\infty} \sum_{k=1}^{r_n} p_{nk} = q$, $0 \leq q < \infty$, and $\lim_{n\to\infty} \max_{1\leq k\leq r_n} p_{nk} = 0$, then*

$$\lim_{n\to\infty} P\{\sum_{k=1}^{r_n} Z_{nk} = i\} = \frac{e^{-q} q^i}{i!}, \quad i = 0, 1, 2, \ldots.$$

Proof for the above lemma can be found in Billingsley (1986). Also note that, by letting $q \to \infty$, $\sum_{k=1}^{r_n} Z_{nk} \to \infty$ in probability.

Proof of Proposition 1: For a small $\epsilon > 0$, let $C_\epsilon = \sup\{|t| : |\kappa_o(t,0)| > \epsilon\}$. Then, as $n \to \infty$,

$$
\begin{aligned}
\sigma_n^{-p} P\{|\kappa_o(X_j/\sigma_n, X/\sigma_n)| > \epsilon\} &= \sigma_n^{-p} P\{\|X_j - X\| < \sigma_n C_\epsilon\} \\
&= \sigma_n^{-p} \int_{\|t-u\| < \sigma_n C_\epsilon} p(t) p(u) dt du \approx v C_\epsilon^p,
\end{aligned}
\tag{31}
$$

where v is the volume of a unit ball in R^p. Let $q(\epsilon) = v C_\epsilon^p$ and let

$$\mathcal{I}_{nj} = \mathcal{I}\{\sigma_n^p |\kappa_n(X_j, X)| > \epsilon\},$$

where $\mathcal{I}$ is an indicator function. Let $S_n = \sum_{j=1}^n \mathcal{I}_{nj}$. Note that from (31) and the definition of r, we have $\sum_{j=1}^n P\{\mathcal{I}_{nj} = 1\} \to rq_\epsilon$. Also note that

$$\max_{1\leq j\leq n} P\{\mathcal{I}_{nj} = 1\} = P\{|\kappa_o(X_j/\sigma_n, X/\sigma_n)| > \epsilon\} \to 0, \quad \text{as } \sigma_n \to 0.$$

Thus, by Lemma 1, for the case $0 \leq r < \infty$

$$P\{\text{card}\{1 \leq j \leq n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} = i\} = P\{S_n = i\} \to \frac{e^{-rq}(rq)^i}{i!}.$$

Also by Lemma 1, for the case $r = \infty$, $\text{card}\{1 \leq j \leq n : \sigma_n^p |\kappa_n(X_j, X)| > \epsilon\} \to \infty$ in probability. $\qquad\square$

Proof of Theorem 8: (1) Let $\tau_n^2 = E\sigma_n^p \kappa_n(X, X)$. Since the fourth moment of $\sigma_n^p \kappa_n(X, X)$ can be bounded by $\sup_t \kappa_o^4(t, t)$, then for any $\eta > 0$

$$\sum_n P\left\{ |n^{-1} \sum_{j=1}^n \sigma_n^p \kappa_n(X_j, X_j) - \tau_n^2| > \eta \right\} \leq \sum_n \frac{\sup_t \kappa_o^4(t,t)}{n^3 \eta^4} < \infty.$$

As $\tau_n^2 \to \tau^2$, hence condition (25) holds almost surely. Let $\nu_n(\epsilon) = P\{|\sigma_n^p \kappa_n(X, X) - \tau_n^2| > \epsilon\}$. Similarly, for any $\epsilon > 0$ and $\eta > 0$

$$\sum_n P\left\{ \left| n^{-1} \sum_{j=1}^n \mathcal{I}\{|\sigma_n^p \kappa_n(X_j, X_j) - \tau_n^2| > \epsilon\} - \nu_n \right| > \eta \right\} < \infty,$$

where $\mathcal{I}$ is an indicator function. As $\nu_n(\epsilon) \to 0$ for all $\epsilon > 0$, condition (26) holds almost surely.

Next, to show conditions (28) and (29) valid almost surely it suffices to show that $\sigma_n^p \kappa_{..} \to 0$ and $\sigma_n^p \kappa_{j.} \to 0$ almost surely. Let C_ϵ be as defined in the proof of Proposition 1. Since that

$$E|\kappa_o(X_j/\sigma_n, X_{j'}/\sigma_n)| \leq \int_{\|x-t\| \leq \sigma_n C_\epsilon} |\kappa_o(x/\sigma_n, t/\sigma_n)| dP(x) dP(t) + \epsilon \to 0, \text{ as } \epsilon \to 0,$$

and that $\kappa_o(X_j/\sigma_n, X_{j'}/\sigma_n)$ is uniformly bounded, then $\sigma_n^p \kappa_{..} \to 0$ and $\sigma_n^p \kappa_{j.} \to 0$ almost surely.

(2) The goal is to show that $n^{-2} \sigma_n^{2p} \sum_{j,j'=1}^n \langle \tilde{\Gamma}(X_j), \tilde{\Gamma}(X_{j'}) \rangle_{\mathcal{H}_n}^2 \to 0$ almost surely. The proof can be resolved into steps below.

- Some straightforward calculation leads to the following decomposition:

$$n^{-2} \sigma_n^{2p} \sum_{j,j'=1}^n \langle \tilde{\Gamma}(X_j), \tilde{\Gamma}(X_{j'}) \rangle_{\mathcal{H}_n}^2$$

$$= \sigma_n^{2p} \langle \bar{\Gamma} - m_n, \bar{\Gamma} - m_n \rangle_{\mathcal{H}_n}^2 - 2n^{-1} \sigma_n^{2p} \sum_{j=1}^n \langle \Gamma(X_j) - m_n, \bar{\Gamma} - m_n \rangle_{\mathcal{H}_n}^2$$

$$+ n^{-2} \sigma_n^{2p} \sum_{j,j'=1}^n \langle \Gamma(X_j) - m_n, \Gamma(X_{j'}) - m_n \rangle_{\mathcal{H}_n}^2,$$

where $m_n(\cdot) = E\kappa_n(X, \cdot)$.

- It is easy to see that $\sigma_n^p \|\bar{\Gamma} - m_n\|_{\mathcal{H}_n}^2 \to 0$ almost surely and then that

$$n^{-1} \sum_{j=1}^n \sigma_n^{2p} \langle \Gamma(X_j) - m_n, \bar{\Gamma} - m_n \rangle_{\mathcal{H}_n}^2$$

$$\leq n^{-1} \sum_{j=1}^n \sigma_n^p \|\Gamma(X_j) - m_n\|_{\mathcal{H}_n}^2 \; \sigma_n^p \|\bar{\Gamma} - m_n\|_{\mathcal{H}_n}^2 \to 0 \; a.s.$$

- Let $\rho_n = \sigma_n^{2p} E\langle \Gamma(X_j) - m_n, \Gamma(X_{j'}) - m_n \rangle_{\mathcal{H}_n}^2$. Since all the random variables given by $\sigma_n^{2p} \langle \Gamma(X_j) - m_n, \Gamma(X_{j'}) - m_n \rangle_{\mathcal{H}_n}^2 - \rho_n$, where $j \neq j'$, have zero mean and a common bounded variance, then

$$\frac{2}{n(n-1)} \sum_{1 \leq j < j' \leq n} \sigma_n^{2p} \langle \Gamma(X_j) - m_n, \Gamma(X_{j'}) - m_n \rangle_{\mathcal{H}_n}^2 - \rho_n \to 0 \; a.s.$$

- Finally, we show that $\lim_{n\to\infty} \rho_n = 0$. The proof goes as follows.

$$\begin{aligned} \rho_n &= \sigma_n^{2p} E\langle \Gamma(X_j) - m_n, \Gamma(X_{j'}) - m_n \rangle_{\mathcal{H}_n}^2 \\ &\leq \sigma_n^{2p} E(\|\Gamma(X_j) - m_n\|_{\mathcal{H}_n}^2 \|\Gamma(X_{j'}) - m_n\|_{\mathcal{H}_n}^2) \\ &= \sigma_n^{2p} \{E\|\Gamma(X) - m_n\|_{\mathcal{H}_n}^2\}^2 = \sigma_n^{2p} \{E\|\Gamma(X)\|_{\mathcal{H}_n}^2 - \|m_n\|_{\mathcal{H}_n}^2\}^2 \\ &= \{var(\kappa_o(X/\sigma_n, X/\sigma_n))\}^2 \to 0. \end{aligned}$$

$\square$

References

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* 2nd ed., Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686, 337–404.

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12, 2385–2404.

Billingsley, P. (1986). *Probability and Measure.* 2nd ed., John Wiley & Sons, New York.

Chen, C.H. and Li, K.C. (2001). Generalization of Fisher's linear discriminant analysis via the approach of sliced inverse regression. *J. Korean Statist. Soc.*, 30, 193–217.

Cheng, Y.J., Mao, S.L., Guo, H.M., Fang, T.S., Lin, J.H., and Lin, J.T. (1994). Study of indictors of science education: learning progress. *NSC Report No. 83-0111-S001-001*, National Science Council, Taiwan, ROC.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, 273–279.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12, 793–815.

Glass, G.V. (1978). Standards and criteria. *J. Educational Measurement*, 15, 237-261.

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för Matematik*, 1, 195–277.

Grenander, U. (1963). *Probabilities on Algebraic Structures.* Almqvist & Wiksells, Stockholm, and John Wiley & Sons, New York.

Grenander, U. (1981) *Abstract Inference.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc.* B, 58, 155–176.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag, New York.

Hein, M. and Bousquet, O. (2004). Kernels, associated structures and generalizations. Technical report, Max Planck Institute for Biological Cybernetics, Germany. http://www.kyb.tuebingen.mpg.de/techreports.html.

Huang, S.Y. and Lee, Y.J. (2004). Reduced support vector machines: a statistical theory. Technical report, Institute of Statistical Science, Academia Sinica, Taiwan. http://stat.sinica.edu.tw/syhuang.

Janson, S. (1997). *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge.

Lee, Y.J. and Mangasarian, O.L. (2001). RSVM: reduced support vector machines. *Proceeding 1st International Conference on Data Mining*, SIAM.

Lin, M.H., Huang, S.Y. and Chang, Y.C. Kernel-based discriminant techniques for educational placement. *J. Educational & Behavioral Statistics*, 29, 219–241.

Mahalanobis, P.C. (1925). Analysis of race mixture in Bengal. *J. Asiat. Soc.* (Bengal), 23, 301.

Mika, S. (2002). *Kernel Fisher Discriminants*. Ph.D. dissertation, Electrical Engineering and Computer Science, Technische Universität Berlin.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B. and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In Hu, Y.-H., Larsen, J., Wilson, E., and Douglas, S., eds, *Neural Networks for Signal Processing*, IX, 41–48, IEEE.

Mika, S., Rätsch, G. and Müller, K.-R. (2001). A mathematical programming approach to the kernel Fisher Algorithm. In T.K. Leen, T.G. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems*, 13, 591–597, MIT Press.

Mika, S., Smola, A. and Schölkopf, B. (2001). An improved training algorithm for kernel Fisher discriminants. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics*, 98–104, Morgan Kaufmann.

Rao, C.R. and Varadarajan, V.S. (1963). Discrimination of Gaussian processes. *Sankhyā*, A, 25, 303–330.

Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Taxt, T., Hjort, N. and Eikvil, L. (1991). Statistical classification using a linear mixture of multinormal probability densities. *Pattn. Recogn. Lett.*, 12, 731–737.

Vakhania, N.N. Tarieladze, V.I. and Chobanyan, S.A. (1987). *Probability Distributions on Banach Spaces*. Translated from the Russian by W.A. Woyczynski. Mathematics and Its Applications (Soviet Series), 14, D. Reidel Publishing Co., Dordrecht, Holland.

Van Gestel, T., Suykens, J.A.K. and De Brabanter, J. (2001). Least squares support vector machine regression for discriminant analysis. *Proc. International Joint INNS-IEEE Conf. Neural Networks (INNS2001)*, 2445–2450. Wiley, New York.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley, New York.

Xu, J., Zhang, X. and Li, Y. (2001). Kernel MSE algorithm: a unified framework for KFD, LS-SVM and KRR. *Proceedings Intern. Joint Conf. Neural Networks*, 2, 1486–1491, IEEE Press.