**DOI:** http://dx.doi.org/10.1016/j.patrec.2011.04.007

# On the Equivalence of Kernel Fisher Discriminant Analysis and Kernel Quadratic Programming Feature Selection

I. Rodriguez-Lujan[a,*], C. Santa Cruz[a], R. Huerta[b]

[a] *Departamento de Ingeniería Informática and Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, 28049 Madrid, Spain.*

[b] *BioCircuits Institute, University of California San Diego, La Jolla CA 92093-0404, USA.*

## Abstract

We reformulate the Quadratic Programming Feature Selection (QPFS) method in a kernel space to obtain a vector which maximizes the quadratic objective function of QPFS. We demonstrate that the vector obtained by Kernel Quadratic Programming Feature Selection is equivalent to the Kernel Fisher vector and, therefore, a new interpretation of the Kernel Fisher Discriminant Analysis is given which provides some computational advantages for highly unbalanced datasets.

*Keywords:* Kernel Fisher Discriminant, Quadratic Programming Feature Selection, Feature Selection, Kernel Methods.

## 1. Introduction

Identifying a proper representation of data is a problem of growing importance in machine learning because of the increasing size and dimensionality of real-world datasets. Linear feature selection and extraction methods, such as Principal Component Analysis (PCA) (Jolliffe, 2002), Canonical Correlation Analysis (CCA) (Afifi and Clark, 1999) and Linear Discriminant Analysis (LDA) (Fukunaga, 1972), are preferable due to their computational speed and simplicity but for most real-world data they are not complex enough. They are conducted in the original space and cannot handle nonlinear relationships in the data. One option to tackle this problem is making use of kernel methods (Shawe-Taylor and Cristianini, 2004) which maps the data from an original space to a *feature space* $\mathcal{F}$ via a (nonlinear) mapping $\Phi : \mathbb{R}^l \longrightarrow \mathcal{F}$. The dot-product in

[*] Corresponding author. Tel: +34 91 497 2339; fax: +34 91 497 2334

*Email addresses:* irene.rodriguez@iic.uam.es (I. Rodriguez-Lujan), carlos.santacruz@iic.uam.es (C. Santa Cruz), rhuerta@ucsd.edu (R. Huerta)

the feature space $\mathcal{F}$ is defined by a Mercer kernel (Mercer, 1909) $K : \mathbb{R}^l \times \mathbb{R}^l \longrightarrow \mathbb{R}$ and, the reformulation of traditional linear methods using only dot-products of training samples yields implicitly a nonlinear method in the input space. Examples of these methods are Kernel-PCA (Schlkopf et al., 1998), Kernel-CCA (Lai and Fyfe, 2000) and the Kernel Fisher Discriminant (Mika et al., 1999).

In this work, we adapt our previous feature selection method QPFS (Rodriguez-Lujan et al., 2010) in a kernel space to provide a vector in the kernel space which maximizes the quadratic objective function. Using the Quadratic Program representation of the KFD proposed by (Mika et al., 2000), we demonstrate the equivalence between KFD and KQPFS. This equivalence provides a new interpretation of the Kernel Fisher vector which only depends on the kernel matrix and the labels of training samples making unnecessary the kernelized between and within class scatter matrices calculation. We also study the training cost of both algorithms.

The present manuscript is organized as follows. Section 2 reformulates the Kernel Fisher Discriminant Analysis to a Quadratic Program. Section 3 presents the formulation of the QPFS algorithm in a kernel space, including a regularized version to overcome numerical problems. Section 4 shows the equivalence between KFD and KQPFS and how this equivalence provides a new interpretation of KFD. Section 5 compares their computational complexity. Finally, Section 6 shows the empirical equivalence of KFD and KQPFS in several well-known artificial and real-world datasets. The runtime of both methods as a function of the class label prior probabilities is also provided.

## 2. Kernel Fisher Discriminant

Let $\mathcal{X}_1 = \{x_1^1, \ldots, x_{l_1}^1\}$ and $\mathcal{X}_2 = \{x_1^2, \ldots, x_{l_2}^2\}$ be samples from two different classes, $x_i \in \mathbb{R}^d$ and $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ the complete set of $l$ $(l = l_1 + l_2)$training samples. And let $y \in \{-1, 1\}^l$ be the vector with the corresponding labels.

The Kernel Fisher Discriminant (KFD) consists on finding nonlinear directions by first mapping the data nonlinearly into the feature space $\mathcal{F}$ and computing Fisher's linear discriminant there (Mika et al., 1999).

Specifically, let $\Phi : \mathbb{R}^d \longrightarrow \mathcal{F}$ be the mapping function to the kernel space and $K(x, y) =< \Phi(x), \Phi(y) >$ the Mercer kernel which defines the dot-product in $\mathcal{F}$. To find the linear discriminant in $\mathcal{F}$ we need to maximizing,

$$J(w) = \frac{w^T S_B^\Phi w}{w^T S_W^\Phi w} \tag{1}$$

where $w \in \mathcal{F}$ and $S_B^\Phi$ and $S_W^\Phi$ are the corresponding between and within scatter matrices in $\mathcal{F}$, i.e.

$$\begin{aligned} S_B^\Phi &= (m_1^\Phi - m_2^\Phi)(m_1^\Phi - m_2^\Phi)^T \\ S_W^\Phi &= \sum_{i=1,2} \sum_{x \in \chi_i} (\Phi(x) - m_i^\Phi)(\Phi(x) - m_i^\Phi)^T \end{aligned}$$

with $m_i^\Phi = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j^i)$.

Finding a solution to Equation 1 in the kernel space $\mathcal{F}$ requires to reformulate it in terms of only dot products of the input patterns (Mika et al., 1999). From the theory of reproducing kernels (Saitoh, 1988), any solution $w \in \mathcal{F}$ must lie in the span of all training samples in $\mathcal{F}$. Therefore $w$ can be expressed as,

$$w = \sum_{i=1}^{l} \alpha_i \Phi\left(x_i\right)$$

Therefore, maximizing Equation 1 is equivalent to maximize

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}$$

being

$$
\begin{aligned}
M &= (M_1 - M_2)(M_1 - M_2)^T \\
(M_i)_j &= \frac{1}{l_i}\sum_{k=1}^{l_i} K(x_j, x_k^i) \\
N &= \sum_{j=1,2} K_j(I - 1_{l_j})K_j^T
\end{aligned}
$$

where $K_j$ is a $N \times l_j$ matrix with $(K_j)_{nm} = k(x_n, x_m^j)$, $I$ is the identity matrix and $1_{l_j}$ the matrix with all entries $\frac{1}{l_j}$.

This problem can be solved by finding the leading eigenvector of $N^{-1}M$ or computing $\alpha^*_{\text{KFD}} = N^{-1}(M_2 - M_1)$. In the last case, some kind of regularization is needed because the problem is ill-posed (Tikhonov and Arsenin, 1977): the dimension of the feature space is usually larger than the number of training samples, which makes matrix $N$ not positive. Regularization functions as $\|\alpha\|^2$, $\|w\|^2$ and others have been proposed in (Mika et al., 1999; Friedman, 1989; Hastie et al., 1993). In (Mika et al., 1999), the matrix $N$ is approximated by $N_\mu = N + \mu_N I$ being $\mu_N$ the minimum value which makes $N$ positive definite.

## 3. Kernel Quadratic Programming Feature Selection

The proposed QPFS method (Rodriguez-Lujan et al., 2010) consists on minimizing a multivariate quadratic function subjected to linear constraints as follows,

$$
\begin{aligned}
\min_x \quad & \tfrac{1}{2}x^T Q x - F^T x \\
\text{s.t.} \quad & x_i \geq 0 \qquad \forall i = 1 \ldots M \\
& \|x\|_1 = 1.
\end{aligned} \tag{2}
$$

Where $x$ is an $d$ -dimensional vector, $Q \in \mathbb{R}^{d \times d}$ is a symmetric positive semidefinite matrix, and $F$ is a vector in $\mathbb{R}^d$ with non-negative entries. $Q$ represents the similarity among variables (redundancy), and $F$ measures how correlated each feature is with the target class (relevance). The components of the solution vector $x^*$ represents the weight of each feature, and we chose to normalize the contribution of each feature to the cost function. Thus, the aim of Equation 3 is to select those features which provide a good tradeoff between relevance and redundancy for the classification task.

The formulation of Equation 3 in a kernel space is not straightforward. For some kernels, it is not possible to give a weight to each feature in the kernel space due to its potential infinite dimension. However, maintaining the goal of redundancy minimization of the features and relevance maximization of each feature with the target class, the Equation 3 can be adapted to find an optimal direction $w$ to project the data into the kernel space. As before, let $\Phi$ be the nonlinear mapping to the feature space $\mathcal{F}$ then, the adapted QPFS objective function is defined as,

$$\min_{x} \frac{1}{2} w^T Q^{\Phi} w - \left(F^{\Phi}\right)^T w \tag{3}$$

where $Q^{\Phi}$ is the redundancy among features in the kernel space, $F^{\Phi}$ is the relevance of each feature with the target class in the kernel space. Thus Equation 3 represents a feature extraction method, KQPFS, instead of a feature selection technique as the original QPFS method.

In the original QPFS approach, correlation and mutual information were considered as similarity measures of redundancy and relevance. For our problem in Equation 3, a linear dependence must be applied because $w$ induces a linear projection of the data. Intuitively, it is possible to adapt mutual information or correlation in the kernel space. However, the mapping function $\Phi$ is usually *implicit* and the dimension of the kernel space $\mathcal{F}$ may be infinite forcing the search of a basis set in the kernel space. If instead of mutual information or correlation, the covariance is used as similarity measure, the KQPFS formulation does not require the presence of an explicit basis in the kernel space. More precisely, the $Q^{\Phi}$ and $F^{\Phi}$ matrices are defined as follows,

$$\begin{aligned} Q^{\Phi} &= \sum_{x \in \mathcal{X}} \left(\Phi(x) - m^{\Phi}\right) \left(\Phi(x) - m^{\Phi}\right)^T \\ F^{\Phi} &= \sum_{x \in \mathcal{X}} \left(y_x - m^y\right) \left(\Phi(x) - m^{\Phi}\right) \end{aligned}$$

where $m^{\Phi}$ and $m^y$ are the mean value of the training samples and the training labels, respectively. That is,

$$
\begin{aligned}
m^{\Phi} &= \frac{1}{l}\sum_{x\in\mathcal{X}}\Phi(x) \\
m^{y} &= \frac{1}{l}\sum_{i=1}^{l} y_i \,.
\end{aligned}
$$

Again, we first need a formulation of Equation 3 in terms of only dot products of input patterns and applying the theory of reproducing kernels (Saitoh, 1988), $w$ is represented as $w = \sum_{i=1}^{l}\alpha_i\Phi\left(x_i\right)$. Therefore, Equation 3 can be formulated as the minimization of function $G(\alpha)$,

$$
G(\alpha) = \frac{1}{2}\alpha^{T}K\left(I - 1_l\right)K\alpha - y^{T}\left(I - 1_l\right)K\alpha \tag{4}
$$

where $I$ is the $l$-dimensional identity matrix and $1_l$ is a $l$-dimensional square matrix with all entries $\frac{1}{l}$.

Let $Q_K = K\left(I - 1_l\right)K$ and $F_K = K\left(I - 1_l\right)y$, the optimal value of $\alpha^{*}_{\text{KQPFS}}$ is obtained making the gradient of $G(\alpha)$ equals to zero,

$$
\alpha^{*}_{\text{KQPFS}} = \left(Q_K\right)^{-1}F_K
$$

If the $Q_K$ matrix is invertible, the formulation of the optimal direction is straightforward,

$$
\begin{aligned}
\alpha^{*}_{\text{KQPFS}} &= \left(Q_K\right)^{-1}F_K \\
&= K^{-1}\left(I - 1_l\right)^{-1}K^{-1}K\left(I - 1_l\right)y \\
&= K^{-1}y
\end{aligned}
$$

Unfortunately, the matrix $Q_K = K\left(I - 1_l\right)K$ is always singular because its rank is upper-bounded by the rank $l-1$ of matrix $\left(I - 1_l\right)$. Therefore, following (Mika et al., 1999), a multiple of the identity matrix is added to $Q_K$ matrix: $Q_\mu = Q_K + \mu_Q I$.

Replacing $Q_K$ by $Q_\mu$ in Equation 4, we obtain the regularized version of KQPFS,

$$
G_\mu(\alpha) = \frac{1}{2}\alpha^{T}\left(Q_K + \mu_Q I\right)\alpha - F_K^{T}\alpha
$$

which is equivalent to,

$$G_\mu(\alpha) = \frac{1}{2}\alpha^T Q_K \alpha - F_K^T \alpha + \frac{\mu_Q}{2}\|\alpha\|^2. \tag{5}$$

And the regularized KQPFS direction is given by,

$$\alpha^*_{\text{KQPFS}} = (Q_K + \mu_Q I)^{-1} F_K \tag{6}$$

$\mu_Q$ is the minimum value which makes $Q_\mu$ positive definite. A process to estimate the parameter $\mu_Q$ is needed. The KQPFS solution obtained in Equation 6 has an easy interpretation as the projection direction which minimizes the covariance among features in the kernel space and maximizes the covariance of each feature in the kernel space with the target class. Moreover, the expression of such direction is quite simple depending only on the kernel matrix $K$ and the class labels $y$.

## 4. Equivalence of KFD and KQPFS

In this section we will demonstrate that the optimal solution of KQPFS is equivalent to the solution of KFD when the same regularization criteria is applied in both cases. Without loss of generality, we will use the regularization defined in Sections 2 and 3. It is straightforward to show that the following proof is also valid for other regularization functions.

As shown in (Mika et al., 2000), the KFD can be reformulated as the following quadratic programming problem,

$$\min_\alpha \alpha^T N \alpha + CP(\alpha) \tag{7}$$

Subject to:

$$\alpha^T (M_1 - M_2) \quad = \quad 2 \tag{8}$$

where $P(\alpha)$ is a regularization term which makes explicit the $N$ regularization and $C \in \mathbb{R}$ the regularization constant. It can be shown (Mika et al., 2000) that solving the problem given in Equations 7 and 8 is equivalent to optimize,

$$\min_{\alpha,b,\xi} \|\xi\|^2 + CP(\alpha) \tag{9}$$

Subject to:

$$\begin{aligned} K\alpha + \overrightarrow{1}b &= y + \xi & (10)\\ \overrightarrow{1}_i^T\xi &= 0 \text{ for } i = 1,2 & (11) \end{aligned}$$

being $\overrightarrow{1} \in \mathbb{R}^l$ a vector with all entries 1 and $\overrightarrow{1}_i R^l$ binary vectors with j-th entry equals to 1 if the j-th sample belongs to class $i$ and 0 otherwise. The quadratic optimization problem defined in Equations 9-11 can be understood as the minimization of the variance of the data along the projection and the maximization of the distance between the average outputs for each class at the same time.

Replacing $N$ by $N_\mu$ in Equation 7, the regularization term $P(\alpha)$ is equal to $\|\alpha\|^2$, the regularization constant $C$ is $\mu_N$ and the regularized quadratic problem in Equations 9-11 is reformulated as,

$$\min_{\alpha,b,\xi} \|\xi\|^2 + \mu_N \|\alpha\|^2. \tag{12}$$

Subject to:

$$\begin{aligned} K\alpha + \overrightarrow{1}b &= y + \xi & (13)\\ \overrightarrow{1}_i^T\xi &= 0 \text{ for } i = 1,2 & (14) \end{aligned}$$

**Proposition 1.** *Given $\mu_N \in \mathbb{R}$ and let $\mu_N = \mu_Q$, any optimal solution $(\alpha^*, b^*, \xi^*)$ to the optimization problem (12-14) is also optimal for (5) and vice versa.*

Proof. Working out $\xi$ in the constraint given in Equation 14 leads to

$$\xi(\alpha, b) = K\alpha + \overrightarrow{1}b - y\ .$$

By expanding $\|\xi(\alpha, b)\|^2$ the optimization problem of Equation 12 is reformulated as

$$\min_{\alpha,b}\{\alpha^T KK\alpha - lb^2 - 2y^T K\alpha + y^T y + \mu_N \|\alpha\|^2\}$$

subject to:

$$\overrightarrow{1}_i^T \xi(\alpha, b) = 0 \text{ for } i = 1, 2$$

The value of $b$ can be expressed as a function of $\alpha$ using the second constraint:

$$b(\alpha) = -\frac{1}{l} 1_l K\alpha + 1_l y \; . \tag{15}$$

Therefore, we have an optimization problem with no constraints:

$$\begin{aligned} \min_\alpha \quad & \{\alpha^T KK\alpha - l\,(b(\alpha))^2 & (16) \\ & -2y^T K\alpha + y^T y + \mu_N \|\alpha\|^2\}. & (17) \end{aligned}$$

Then, substituting $b(\alpha)$ in Equation 17 by the value obtained in Equation 15 we obtain

$$\begin{aligned} \min_\alpha \quad & \{\alpha^T K\,(I - 1_l)\,K\alpha \\ & -2y^T\,(I - 1_l)\,K\alpha + \frac{\mu_N}{2}\|\alpha\|^2 + D\} & (18) \end{aligned}$$

with $D$ being a constant. It follows that the minimum value of Equation 18 is the same as the obtained for the objective function of the regularized KQPFS (Equation 5) when $\mu_N = \mu_Q$.

□

This equivalence provides a new solution of the Fisher direction which not depends explicitly on the un-intuitive kernelized within scatter matrix $N$ (Equation 6). Moreover, the Fisher solution has a simple interpretation as the direction which minimizes the covariance among features and maximizes the covariance of each feature with the target class.

## 5. Computational Cost Comparison

In this section we study the computational cost of KFD and KQPFS to determine whether it is possible to get any computational advantage from the new KFD formulation as the kernelization of QPFS. Even though several algorithms have been proposed to speed up KFD (Cai, 2007; Mika, 2001; Xiong et al., 2004) we are interested in analyzing an equivalent problem to the KQPFS as given in Equation 6. Let us to obtain the *standard* KFD solution as $\alpha^*_{\text{KFD}} = (N_\mu)^{-1}(M_1 - M_2)$ where matrices $N_\mu$, $M_1$ and $M_2$ are defined in

```
1: INPUT: $l, K, y, \mu_N$
2: pos = (y==1);
3: neg = (y==-1);
4: l1 = sum(pos);
5: l2= sum(neg);
6: N =
7: K(:,pos)*(eye(l1)-
   (1/l1)*ones(l1))*(K(:,pos))'+
8: K(:,neg)*(eye(l2)-
   (1/l2)*ones(l2))*(K(:,neg))'+
9: diag($\mu_N$*ones(l,1));
10: M = ((1/l1)*(sum(K(:,pos),2))) -
11: ((1/l2)*(sum(K(:,neg),2)));
12: $\alpha_{\mathrm{KFD}}$ = N\M;
13: OUTPUT: $\alpha_{\mathrm{KFD}}$
```

```
1: INPUT: $l, K, y, \mu_Q$
2: A=K*(eye(l)-((1/l)*ones(l));
3: Q=A*K+diag($\mu_Q$*ones(l,1));
4: B = A*y;
5: $\alpha_{\mathrm{KQPFS}} = Q \setminus B$
6: OUTPUT: $\alpha_{\mathrm{KQPFS}}$
```

Figure 1: MATLAB code of KFD (left) and KQPFS (right) algorithms.

Section 2. Figure 1 shows the MATLAB code for both methods. The number of float-point operations needed by KFD is $4l$ (lines 2-5), $l_1^2 + l_2^2 + l^2 + 2l(l_1^2 + l_2^2) + 3l^2$ (lines 6-9), $l^2 + 3l$ (lines 10-11) and $O(l^3)$ (line 12) which makes a total cost of $O(l^3) + 2l(l_1^2 + l_2^2) + 5l^2 + l_1^2 + l_2^2 + 7l$ operations. In the case of the KQPFS algorithm, $l^2 + l^3$ operations are needed in line 2, $2l^2 + l^3$ in line 3, $l^2$ in line 4 and $O(l^3)$ in line 5 that is, a total cost of $O(l^3) + 2l^3 + 4l^2$ float-point operations. As the line 12 of KFD and line 5 of KQPFS work with dimensionality equivalent matrices, we will suppose that the cost of these lines is the same in both cases therefore, we obtain that KQPFS is computationally faster than the proposed version of KFD if $(l_1^2 + l_2^2)(2l + 1) + 5l^2 + 7l \gg 2l^3 + 4l^2$. The inequality is satisfied when the prior distributions of the class labels are highly unbalanced i.e., when $l_1 \to l$ or $l_2 \to l$. Summing up, the KFD cost depends on the prior distribution of classes and KQPFS is more efficient for highly unbalanced classification problems.

## 6. Experiments

A theoretical proof of the equivalence between KFD and KQPFS has been given in Proposition 1 and in this section we show that the numerical solutions given by KFD and KQPFS provide the same projection direction.

We followed part of the experimental setup described in (Mika et al., 1999): for KFD and KQPFS we used Gaussian kernels and the regularized matrices $N_\mu$ and $Q_\mu$ as described in Sections 2 and 3, respectively.

Thirteen artificial and real world datasets were considered from the Rätsch benchmark repository[1]. Some of these datasets were not binary so they were transformed into two-classes problems and all of them were partitioned into 100 pairs of training and test sets (about 60%:40%).

The experiments require to estimate two parameters, the width of the Gaussian kernel $K(x, y) = e^{\frac{\|x-y\|^2}{\sigma}}$ and the regularization parameter $\mu_N$ of the within class scatter matrix $N$ in KFD (see section 2). The procedure to estimate these parameters consists on running 5-fold cross validation on the first five realizations of the training sets and taking the model parameter to be the median over the five estimates. The value of these parameters is known (Mika et al., 1999). Note that the equivalence of KQPFS and KFD holds when the same regularization form and regularization constant is applied in both cases. Therefore, there is no need to estimate the KQPFS regularization parameter $\mu_Q$.

The empirical equivalence of KFD and KQPFS has been confirmed measuring the cosine between the solutions $\alpha^*_{\text{KFD}}$ and $\alpha^*_{\text{KQPFS}}$. Ideally, the value of the cosine should be close to 1 or to $-1$ which means parallel directions. In all the datasets, the cosine of both directions was 1 for every training set.

Finally, let us provide numerical results of the KFD and KQPFS complexity analyzed in Section 5. The experiment consists on modifying the prior probability of one of the classes, without loss of generality the class of positive labels, and compare the runtime of KFD and KQPFS codes (Figure 1). The regression dataset Abalone available in the LIBSVM repository (Chang and Lin, 2001) was used. The dataset has 4177 samples ($l$) in a 8-dimensional space. To carry out the experiments, the samples were arranged in ascending order according to the regression variable and the prior probability of the positive class $p_1$ was modified from 0 to 1 with a stepwise of 0.05. A pattern is assigned to the positive class if it is among the first $p_1 l$ patterns. Figure 2 shows the runtime in training as a function of the prior probability of the positive class. As expected, the KFD algorithm cost is dependent on the class prior probabilities being faster than KQPFS except when the class distributions are highly unbalanced. The KQPFS complexity is independent on the prior distributions.

[1] The datasets are available at `http://ftp.tuebingen.mpg.de/pub/fml/raetsch-lab/benchmarks/`

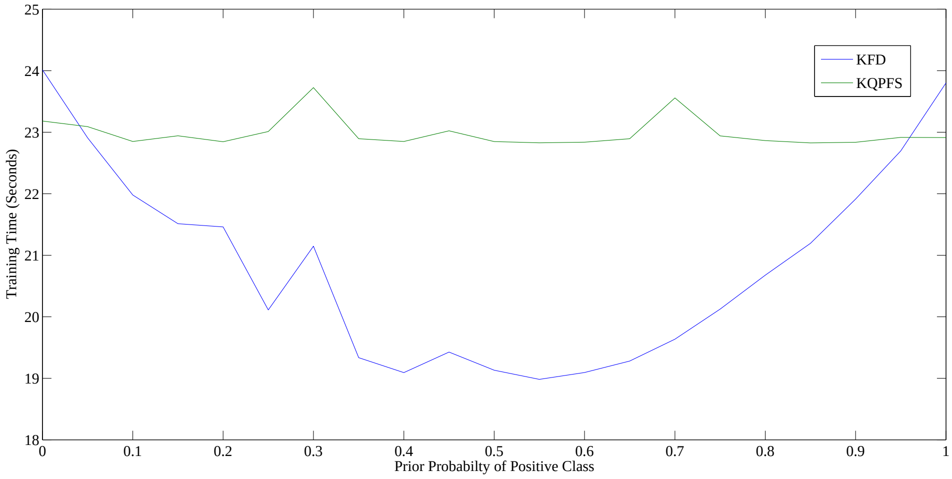

Figure 2: Abalone. Training time in seconds for the KFD and KQPFS algorithms.

## 7. Conclusions

This paper reformulates the Quadratic Programming Feature Selection (QPFS) method to obtain an optimal projection direction in a kernel space (KQPFS). The projection direction given by KQPFS is equivalent to those obtained by the Kernel Fisher Discriminant (KFD) which leads to a new interpretation of the KFD vector as the direction which minimizes the covariance among features and maximizes the covariance of each feature with the target class in the kernel space. This equivalence provides a new solution for KFD disregarding the explicitly dependence on the kernelized between and within scatter matrices. In addition, a more efficient computation of the Kernel Fisher direction is proposed when the classes are highly unbalanced.

## Acknowledgments

I.R.-L. is supported by an FPU grant from Universidad Autónoma de Madrid, and partially supported by the Universidad Autónoma de Madrid-IIC Chair. R.H. acknowledges partial support by ONR N00014-07-1-0741, NIDCD 1R01DC011422-01 and JPL-1396686.

## References

Afifi, A. A., Clark, V., 1999. Computer-Aided Multivariate Analysis, 2nd Edition. Chapman & Hall, Ltd., London, UK, UK.

Cai, D., 2007. Efficient kernel discriminant analysis via spectral regression. Tech. rep.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Friedman, J. H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84 (405), pp. 165–175.

Fukunaga, K., 1972. Introduction to Statistical Pattern Recognition. New York, San Francisco, London: Academic Press.

Hastie, T. J., Tibshirani, R. J., Buja, A., 1993. Flexible discriminant analysis by optimal scoring. Tech. rep., AT&T Bell Laboratories.

Jolliffe, I. T., 2002. Principal Component Analysis. Springer, New York, NY, USA.

Lai, P. L., Fyfe, C., 2000. Kernel and nonlinear canonical correlation analysis. Int. J. Neural Syst. 10 (5), 365–377.

Mercer, J., 1909. Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society, London 209, 415–446.

Mika, S., 2001. An improved training algorithm for kernel fisher discriminants. In: In Proceedings AISTATS 2001. Morgan Kaufmann, pp. 98–104.

Mika, S., Rtsch, G., Mller, K.-R., 2000. A mathematical programming approach to the kernel fisher algorithm. In: Leen, T. K., Dietterich, T. G., Tresp, V. (Eds.), NIPS. MIT Press, pp. 591–597.

Mika, S., Rtsch, G., Weston, J., Schlkopf, B., Mller, K.-R., 1999. Fisher discriminant analysis with kernels.

Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C. S., August 2010. Quadratic programming feature selection. J. Mach. Learn. Res. 99, 1491–1516.

Saitoh, S., 1988. Theory of Reproducing Kernels and its Applications. Longman Scientific&Technical, Harlow, England.

Schlkopf, B., Smola, A., Mller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10 (5), 1299–1319.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK.

Tikhonov, A. N., Arsenin, V. Y., 1977. Solutions of Ill-posed problems. W.H. Winston.

Xiong, T., Ye, J., Li, Q., Janardan, R., Cherkassky, V., 2004. Efficient kernel discriminant analysis via qr decomposition. In: NIPS.