
Matching Pursuit Kernel Fisher Discriminant Analysis

Tom Diethe, Zakria Hussain, David R. Hardoon and John Shawe-Taylor

Centre for Computational Statistics and Machine Learning

Department of Computer Science

University College London, UK, WC1E 6BT

{ t.diethe,z.hussain,d.hardoon,j.shawe-taylor } @ cs.ucl.ac.uk

Abstract

We derive a novel sparse version of Kernel Fisher Discriminant Analysis (KFDA) using an approach based on Matching Pursuit (MP). We call this algorithm Matching Pursuit Kernel Fisher Discriminant Analysis (MPKFDA). We provide generalisation error bounds analogous to those constructed for the Robust Minimax algorithm together with a sample compression bounding technique. We present experimental results on real world datasets, which show that MPKFDA is competitive with the KFDA and the SVM on UCI datasets, and additional experiments that show that the MPKFDA on average outperforms KFDA and SVM in extremely high dimensional settings.

1 INTRODUCTION

Fisher Discriminant Analysis (FDA) was proposed by Fisher [1936] as a statistical approach for classifying new data into two separate groups. In some sense, FDA can be viewed as a classifier coupled with the ability to carry out dimensionality reduction similar to Principal Components Analysis (PCA). Fisher's Discriminant has been formulated using the kernel trick, resulting in Kernel Fisher Discriminant Analysis (KFDA) [Mika et al., 1999]. One drawback, as with most kernel methods, is that storing large kernel matrices is computationally prohibitive. In order to tackle this problem, several authors have made attempts at addressing this issue by creating low rank kernel matrices behaving similarly to the full ranked

ones [Bach and Jordan, 2005, Kulis et al., 2006]. Most importantly for us is the work of Smola and Schölkopf [2000] where they devise a method of constructing low rank kernel matrices, motivated by a greedy approach called Matching Pursuit.

Matching Pursuit (MP) was proposed in the signal processing literature by Mallat and Zhang [1993] as an attempt at finding a sparse number of dictionaries from a signal. In many ways this problem can be interpreted as a sparse version of least squares regression when the Orthogonal Matching Pursuit (OMP) [Pati et al., 1993] version is applied. In OMP each time a dictionary is chosen the remaining weight vectors are projected into a space orthogonal to those chosen such that future dictionaries are only considered from a set far from those already picked. As with (K)FDA, Kernel Matching Pursuit (KMP) [Vincent and Bengio, 2002] has been proposed as MP's kernel counterpart.

We take the idea of Matching Pursuit, due to its very fast greedy iterative nature, and apply it to Kernel Fisher Discriminant Analysis in order to impose dual sparsity. We prove that this sparse version results in generalisation error bounds guaranteeing its future success. The novel bounds come from the Shawe-Taylor and Cristianini [2003] analysis of the Robust Minimax algorithm of Lanckriet et al. [2003], which is similar in flavour to FDA. Together with the bounds of Shawe-Taylor and Cristianini [2003] we also apply a compression argument [Littlestone and Warmuth, 1986] in order to gain an advantage of the dual sparsity we get from our algorithm. However, the algorithm does not form a traditional compression scheme so we use a similar idea to that of Hussain and Shawe-Taylor [2008] and bound the generalisation error in the sparsely defined subspace by amalgamating both theories mentioned above. In some ways the bounds justify the choice of our fast iterative greedy strategy, which is not provably optimal [Chen et al., 1998], by guaranteeing that for any random choice of dataset and from any given distribution we will be “probably approxi-

Appearing in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

mately correct” [Valiant, 1984] with our predictions.

One of the practical advantages of Matching Pursuit Kernel Fisher Discriminant Analysis (MPKFDA) lies in the evaluation on test points - only k kernel evaluations are required (where k is the number of basis vectors chosen) compared to m (the number of samples) needed for KFDA. It is also worth stating that MPKFDA like the KFDA has the advantage of delivering conditional probabilities of classification (unlike the SVM). There has been some research suggesting that one cannot estimate conditional probabilities without involving all of the data (see Bartlett and Tewari [2007]) - hence kernel methods cannot deliver this efficiently - but here we do take account of all of the data whilst still having an efficient kernel representation.

Previous work on Sparse (Kernel) Fisher Discriminant Analysis includes [Feng and Shi, 2004] who extend on work by Mika et al. [2001] which approximates the weight vector, by proposing a sparse solution of KFDA for face detection using a convex quadratic programme. Similarly, Dundar et al. [2005], motivated by Computer-Aided Detection systems for identifying structures of interest in medical images, have proposed a sparse variation of FDA using 1-norm minimisation. Finally, Xing et al. [2005] have also proposed a sparse KFDA through approximating the implicit within-class scatter matrix in feature space. Our work is different in that we propose a new variant of sparse KFDA that is motivated by matching pursuit and also a novel generalisation error bound that, to our knowledge, has not previously been done.

The paper has the following layout. In Section 2 we present the notations used throughout the paper while Section 3 discusses the main practical contribution of the paper and presents the MPKFDA algorithm. In Section 4 we build on the theory presented in Shawe-Taylor and Cristianini [2003] and Littlestone and Warmuth [1986], Floyd and Warmuth [1995] to propose a novel generalisation error bound upper bounding the loss of MPKFDA. The experiments are given in Section 5. Finally, we conclude with a discussion in Section 6.

2 PRELIMINARIES

Assume we have a sample S containing examples $\mathbf{x} \in \mathbb{R}^n$ and labels $y \in \{-1, 1\}$. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)'$ be the input vectors stored in matrix \mathbf{X} as row vectors, where $'$ denote the transpose of vectors or matrices. For simplicity we always assume that the examples are already projected into the kernel defined feature space, so that the kernel matrix \mathbf{K} has entries $\mathbf{K}[i, j] = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. In the analysis section we will ex-

PLICITLY denote the feature map $\phi(\mathbf{x})$ for some vector \mathbf{x} . The notation $\mathbf{K}[:, i]$ will denote the i th column of the matrix \mathbf{K} . When given a set of indices $\mathbf{i} = \{i_1, \dots, i_k\}$ (say) then $\mathbf{K}[\mathbf{i}, \mathbf{i}]$ denotes the square matrix defined solely by the index set \mathbf{i} .

For analysis purposes we assume that the training examples are generated i.i.d. according to an unknown but fixed probability distribution that also governs the generation of the test data. Expectation over the training examples (empirical average) is denoted by $\hat{\mathbb{E}}[\cdot]$, while expectation with respect to the underlying distribution is denoted $\mathbb{E}[\cdot]$.

For the sample compression analysis the **compression function** Λ induced by a sample compression learning algorithm A on training set S is the map $\Lambda : S \mapsto \Lambda(S)$ such that the *compression set* $\Lambda(S) \subset S$ is returned by A . A **reconstruction function** Ψ is a mapping from a compression set $\Lambda(S)$ to a set F of functions $\Psi : \Lambda(S) \mapsto F$.

Let $A(S)$ be the function output by learning algorithm A on training set S . Therefore, a sample compression scheme is a reconstruction function Ψ mapping a compression set $\Lambda(S)$ to some set of functions F such that $A(S) = \Psi(\Lambda(S))$. If F is the set of Boolean-valued functions then the sample compression scheme is said to be a classification algorithm.

We define $\hat{\mu}(\mu)$ to be the empirical (true) mean of a sample of m points from the set S projected into a higher dimensional space using ϕ ,

$$\begin{aligned} \mu &= \mathbb{E}[\phi(\mathbf{x})], \\ \hat{\mu} &= \frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i), \end{aligned}$$

and $\hat{\Sigma}(\Sigma)$ its empirical (true) covariance matrix.

3 ALGORITHM

Matching Pursuit can be formalised as a general framework in machine learning, where repeating the following steps of:

1. Function maximisation; *and*
2. Deflation,

can result in Matching Pursuit algorithms for learning tasks other than regression. In this paper we present an application of this general framework to Kernel Fisher Discriminant Analysis (KFDA), resulting in a sparse form of KFDA that we call Matching Pursuit Kernel Fisher Discriminant Analysis (MPKFDA).

We can have a Matching Pursuit algorithm for Fisher Discriminant Analysis (see Shawe-Taylor and Cristianini [2004] for details) in the following way. Initially, we pick one example $\mathbf{i} = \{i_1\}$ and project the remaining training examples into the space defined by \mathbf{i} . We then find the index that maximises the Fisher discriminant analysis (FDA) loss. After which we carry out a deflation of the data matrix \mathbf{X} (or kernel \mathbf{K}) to allow new training examples to be chosen. Finally this give us a set \mathbf{i} of training examples that can be used to compute the final weight vector \mathbf{w} , together with the FDA decision function $f(\mathbf{x}) = \text{sgn}(\mathbf{w}'\mathbf{x} + b)$ where b is the bias and \mathbf{x} an example.

Using the notation from Shawe-Taylor and Cristianini [2004], we have the following maximisation problem for FDA:

$$\mathbf{w} = \max_{\mathbf{w}} \frac{\mathbf{w}'\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w}}{\mathbf{w}'\mathbf{X}'\mathbf{B}\mathbf{X}\mathbf{w}}, \quad (1)$$

where m^+ are the number of positive examples, m^- the number of negative examples and $\mathbf{B} = \mathbf{D} - \mathbf{C}^+ - \mathbf{C}^-$ where \mathbf{D} is a diagonal matrix with entries

$$\mathbf{D}[i, i] = \begin{cases} 2m^-/m & \text{if } y_i = +1 \\ 2m^+/m & \text{if } y_i = -1, \end{cases}$$

and \mathbf{C}^+ and \mathbf{C}^- are given by

$$\mathbf{C}^+[i, j] = \begin{cases} 2m^-/(mm^+) & \text{if } y_i = +1 = y_j \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbf{C}^-[i, j] = \begin{cases} 2m^+/(mm^-) & \text{if } y_i = -1 = y_j \\ 0 & \text{otherwise.} \end{cases}$$

We begin by applying the Nyström method of low-rank approximation of the Gram matrix [Williams and Seeger, 2001]

$$\begin{aligned} \tilde{\mathbf{K}} &= \mathbf{K}[:, \mathbf{i}]\mathbf{K}[\mathbf{i}, \mathbf{i}]^{-1}\mathbf{K}[:, \mathbf{i}]' \\ &= \mathbf{K}[:, \mathbf{i}]\mathbf{R}'\mathbf{R}\mathbf{K}[:, \mathbf{i}]', \end{aligned}$$

where \mathbf{R} is the Cholesky decomposition of $\mathbf{K}[\mathbf{i}, \mathbf{i}]^{-1}$ such that $\mathbf{R}'\mathbf{R} = \mathbf{K}[\mathbf{i}, \mathbf{i}]^{-1}$. However, rather than use the full $[m \times m]$ low rank approximation, it would be preferable to work in the $[k \times k]$ space where $k \ll m$. In order to do this we treat $\mathbf{K}[:, \mathbf{i}]\mathbf{R}'$ as a new input \mathbf{X} in FDA, which in effect means we are projecting into a k -dimensional subspace. Within this space we can view

$$\tilde{\Sigma}_k = \mathbf{R}\mathbf{K}[:, \mathbf{i}]\mathbf{K}[:, \mathbf{i}]\mathbf{R}',$$

as a form of covariance matrix within this space. This trick allows us to perform nonlinear discriminant analysis on a sparse subspace using standard linear FDA.

We can define the following maximisation problem for a dual sparse version of FDA by setting $\mathbf{w} = \mathbf{X}'\mathbf{e}_i$ where \mathbf{e}_i is the i^{th} unit vector of length m , and substituting into the FDA problem described above (ignoring constants) to yield:

$$\begin{aligned} \max_i \rho_i &= \frac{\mathbf{e}_i'\mathbf{X}\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{X}'\mathbf{e}_i}{\mathbf{e}_i'\mathbf{X}\mathbf{X}'\mathbf{B}\mathbf{X}\mathbf{X}'\mathbf{e}_i} \\ &= \frac{\mathbf{K}[:, i]'\mathbf{y}\mathbf{y}'\mathbf{K}[:, i]}{\mathbf{K}[:, i]'\mathbf{B}\mathbf{K}[:, i]} \end{aligned}$$

Maximising the quantity above leads to maximisation of the Fisher Discriminant ratio corresponding to \mathbf{e}_i , and hence a sparse subset of the original KFDDA problem. We would like to find the optimal set of indices \mathbf{i} . We proceed in a greedy manner (Matching Pursuit) in much the same way as Smola and Schölkopf [2000] and Vincent and Bengio [2002]. The procedure involves choosing basis vectors that maximise the Fisher Discriminant ratio iteratively until some pre-specified number of k vectors are chosen.

After finding the best index i we would orthogonalise the matrix \mathbf{K} by setting $\boldsymbol{\tau} = \mathbf{K}[:, i]$, and deflating like so:

$$\mathbf{K} = \left(\mathbf{I} - \frac{\boldsymbol{\tau}\boldsymbol{\tau}'}{\boldsymbol{\tau}'\boldsymbol{\tau}} \right) \mathbf{K}.$$

This deflation ensures that remaining potential basis vectors will be chosen from a space that is orthogonal to those bases already picked¹. After choosing the k training examples, giving $\mathbf{i} = (i_1, \dots, i_k)$, we can define:

$$\mathbf{R}\mathbf{K}[:, \mathbf{i}]'$$

as a new data matrix, where \mathbf{R} is the Cholesky decomposition of $\mathbf{K}[\mathbf{i}, \mathbf{i}]^{-1}$. We then train FDA as in Equation 1 in this new projected space to find a k -dimensional weight vector \mathbf{w}_k . Given the index j of a test point \mathbf{x}_j , and using the train-test kernel on this point $\mathbf{K}[j, \mathbf{i}]$ and its projection $\phi(x_j) = \mathbf{R}\mathbf{K}[j, \mathbf{i}]'$, we can make predictions using the FDA prediction function,

$$f(x_j) = \text{sgn}(\langle \tilde{\mathbf{w}}, \phi(\mathbf{x}_j) \rangle + b) \quad (2)$$

4 GENERALISATION ERROR ANALYSIS

We now construct a generalisation error bound for Matching Pursuit Kernel Fisher Discriminant Analysis by applying the results from Shawe-Taylor and Cristianini [2003] with a compression argument.

¹It is assumed that the vectors of the matrix \mathbf{K} do not form an orthonormal basis

Algorithm 1 Matching Pursuit Kernel Fisher Discriminant Analysis

Input: kernel \mathbf{K} , sparsity parameter $k > 0$, training labels \mathbf{y} .

- 1: calculate matrix \mathbf{B}
- 2: initialise $\mathbf{i} = ()$
- 3: **for** $i = 1$ to k **do**
- 4: set \mathbf{i}_i to index of $\max \frac{\mathbf{K}[:,i]'\mathbf{y}\mathbf{y}'\mathbf{K}[:,i]}{\mathbf{K}[:,i]'\mathbf{B}\mathbf{K}[:,i]}$
- 5: set $\boldsymbol{\tau} = \mathbf{K}[:,\mathbf{i}_i]$ to deflate kernel matrix like so:

$$\mathbf{K} = \left(\mathbf{I} - \frac{\boldsymbol{\tau}\boldsymbol{\tau}'}{\boldsymbol{\tau}'\boldsymbol{\tau}} \right) \mathbf{K}$$

- 6: **end for**
- 7: calculate the projection $\mathbf{R}\mathbf{K}[:,\mathbf{i}]'$ where \mathbf{R} is the Cholesky decomposition of $\mathbf{K}[\mathbf{i},\mathbf{i}]^{-1}$ and $\mathbf{i} = (\mathbf{i}_1, \dots, \mathbf{i}_k)$
- 8: train FDA using Equation 1 in this new projected space to find a sparse weight vector $\tilde{\mathbf{w}}$ and make predictions using Equation 2

Output: final set \mathbf{i} , (sparse) weight vector $\tilde{\mathbf{w}}$, bias term b

4.1 PRELIMINARIES

We need the following two results from [Shawe-Taylor and Cristianini, 2003]. The first bounds the difference between the empirical and true means.

Theorem 4.1 (Bound on the true and empirical means). *Let S be an m sample generated independently at random according to a distribution P . Then with probability at least $1 - \delta$ over the choice of S , we have*

$$\|\hat{\mu} - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\| \leq \frac{R}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{1}{\delta}} \right). \quad (3)$$

Consider the covariance matrix defined as

$$\boldsymbol{\Sigma} = \mathbb{E}[(\phi(\mathbf{x}) - \mu)(\phi(\mathbf{x}) - \mu)'].$$

Let the empirical estimate of this quantity be

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbb{E}}[(\phi(\mathbf{x}) - \hat{\mu})(\phi(\mathbf{x}) - \hat{\mu})'].$$

The following corollary bounds the difference between the empirical and true covariance.

Corollary 4.2 (Bound on the true and empirical covariances). *Let S be an m sample generated independently at random according to a distribution P . Then with probability at least $1 - \delta$ over the choice of S , we have*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F \leq \frac{2R^2}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right), \quad (4)$$

where R is the radius of the ball in the feature space containing the support of the distribution and provided

$$m \geq \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right)^2.$$

The following Lemma is connected with a classification algorithm developed by Lanckriet et al. [2003]. The basis for the approach is the following Lemma.

Lemma 4.3. *Let μ be the mean of a distribution and $\boldsymbol{\Sigma}$ its covariance matrix, $\mathbf{w} \neq 0$, b given, such that $\mathbf{w}'\mu \leq b$ and $\alpha \in [0, 1)$, then if*

$$b - \mathbf{w}'\mu \geq \kappa(\alpha) \sqrt{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$, then

$$P(\mathbf{w}'\phi(\mathbf{x}) \leq b) \geq \alpha$$

In order to provide a true error bound we must bound the difference between this estimate and the value that would have been obtained had the true mean and covariance been used.

4.2 BOUND FOR MATCHING PURSUIT KERNEL FISHER DISCRIMINANT ANALYSIS

We apply the bound above to a subspace defined from a small number $k \ll m$ of basis vectors. Let $\mathbf{i} = (i_1, \dots, i_k)$ be a vector of indices used to form a k -dimensional subspace such as the one defined by Matching Pursuit Kernel Fisher Discriminant Analysis (MPKFDA). We use the notation $S_{\mathbf{i}}$ to denote the samples pointed to by \mathbf{i} . First we give a general bound and then specialise it to the case of Matching Pursuit Kernel Fisher Discriminant Analysis.

Theorem 4.4 (main). *Let S be a sample of m points drawn independently according to a probability distribution P where R is the radius of the ball in the feature space containing the support of the distribution. Let $\hat{\mu}_k$ (μ_k) be the empirical (true) mean of a sample of $m - k$ points from the set $S \setminus S_{\mathbf{i}}$ projected into a k -dimensional space, $\hat{\boldsymbol{\Sigma}}_k$ ($\boldsymbol{\Sigma}_k$) its empirical (true) covariance matrix, $\mathbf{w}_k \neq 0$ with norm 1, and b_k given, such that $\mathbf{w}_k'\mu_k \leq b_k$ and $\alpha \in [0, 1)$. Then with probability $1 - \delta$ over the draw of the random sample, if*

$$b_k - \mathbf{w}_k'\hat{\mu}_k \geq \kappa(\alpha) \sqrt{\mathbf{w}_k'\hat{\boldsymbol{\Sigma}}_k\mathbf{w}_k},$$

then

$$P(\mathbf{w}_k'\phi(\mathbf{x}) - b_k > 0) < 1 - \alpha,$$

where

$$\alpha = \frac{(b_k - \mathbf{w}'_k \hat{\mu}_k - A)^2}{\mathbf{w}'_k \hat{\Sigma}_k \mathbf{w}_k + B + (b_k - \mathbf{w}'_k \hat{\mu}_k - A)^2},$$

such that $\|\hat{\mu}_k - \mu_k\| \leq A$ where

$$A = \frac{R}{\sqrt{m-k}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{2m}{\delta}} \right)$$

and $\|\hat{\Sigma}_k - \Sigma_k\|_F \leq B$ where

$$B = \frac{2R^2}{\sqrt{m-k}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{4m}{\delta}} \right).$$

Proof. (sketch). First we re-arrange $b_k - \mathbf{w}'_k \mu \geq \kappa(\alpha) \sqrt{\mathbf{w}'_k \Sigma \mathbf{w}}$ from Lemma 4.3 in terms of $\kappa(\alpha)$:

$$\kappa(\alpha) = \frac{b_k - \mathbf{w}'_k \mu}{\sqrt{\mathbf{w}'_k \Sigma \mathbf{w}}}. \quad (5)$$

These quantities are in terms of the true means and covariances. In order to achieve an upper bound we need the following sample compressed results for the true and empirical means (Theorem 4.1) and covariances (Corollary 4.2):

$$\|\hat{\mu}_k - \mathbb{E}_{\mathbf{x}}[\hat{\mu}_k(\mathbf{x})]\| \leq A = \frac{R}{\sqrt{m-k}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{2m}{\delta}} \right),$$

and

$$\|\hat{\Sigma}_k - \Sigma_k\|_F \leq B = \frac{2R^2}{\sqrt{m-k}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{4m}{\delta}} \right).$$

Given Equation (5) we can use the empirical quantities for the means and covariances in place of the true quantities. However, in order to derive a genuine upper bound we also need to take into account the upper bounds between the empirical and true means. Including these in the expression above for $\kappa(\alpha)$ by replacing δ with $\delta/2$, to derive a lower bound, we get:

$$\kappa(\alpha) = \frac{b_k - \mathbf{w}'_k \hat{\mu}_{S_k} - A}{\sqrt{\mathbf{w}'_k \hat{\Sigma}_k \mathbf{w}_k + B}}.$$

Finally, making the substitution $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$ and solving for α yields the result. \square

The following Proposition upper bounds the generalisation error of Matching Pursuit Kernel Fisher Discriminant Analysis.

Proposition 4.5. *Let \mathbf{w}_k, b_k , be the (normalised) weight vector and associated threshold returned by the Matching Pursuit Kernel Fisher Discriminant Analysis algorithm (MPKFDA) when presented with a training set S . Furthermore, let $\hat{\Sigma}_k^+$ ($\hat{\Sigma}_k^-$) be the empirical covariance matrices associated with the positive (negative) examples of the $m-k$ training samples from $S \setminus S_k$ projected into a k dimensional space using \mathbf{w}_k . Then with probability at least $1 - \delta$ over the draw of the random training set S of m training examples, the generalisation error ϵ is bounded by*

$$\epsilon \leq \max(1 - \alpha^+, 1 - \alpha^-)$$

where $\alpha^j, j = +, -$ such that

$$\alpha^j = \frac{\left(j(\mathbf{w}'_k \hat{\mu}_{S_k}^j - b_k) - C^j \right)^2}{\mathbf{w}'_k \hat{\Sigma}_k^j \mathbf{w}_k + D^j + \left(j(\mathbf{w}'_k \hat{\mu}_{S_k}^j - b_k) - C^j \right)^2},$$

where

$$C^j = \frac{R}{\sqrt{m^j - k^j}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{4m}{\delta}} \right),$$

and

$$D^j = \frac{2R^2}{\sqrt{m^j - k^j}} \left(2 + \sqrt{k \ln \frac{em}{k} + 2 \ln \frac{8m}{\delta}} \right).$$

Proof. For the negative -1 part of the proof we require $b_k - \mathbf{w}'_k \hat{\mu}_k^- \geq \kappa(\alpha) \sqrt{\mathbf{w}'_k \hat{\Sigma}_k^- \mathbf{w}_k}$ which is a straight forward application of Theorem 4.4 with δ replaced with $\delta/2$. For the positive $+1$ part, observe that we require $-b_k + \mathbf{w}'_k \hat{\mu}_k^+ \geq \kappa(\alpha) \sqrt{\mathbf{w}'_k \hat{\Sigma}_k^+ \mathbf{w}_k}$, hence, a further application of Theorem 4.4 with δ replaced by $\delta/2$ suffices. \square

5 EXPERIMENTS

We present a comparison on 13 benchmark datasets derived from the UCI, DELVE and STATLOG benchmark repositories. We analyse the performance of KFDA, MPKFDA, and SVM using Radial Basis Function (RBF) kernels. The data comes in 100 predefined splits into training and test sets (20 in the case of the image and splice datasets) as described in [Mika et al., 1999]². For each of the datasets we used cross-validation (c.v.) to select the optimal parameters (the

²Available to download from:

<http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm>

RBF kernel width parameter, the C parameter in the SVM, and k the number of iterations in MPKFDA). We used 5-fold c.v. over the first five training datasets with a coarse range of parameter values, selecting the median over the five sets as the optimal value, followed by a similar process using a fine range of parameter values. This way of estimating the parameters leads to more robust comparisons between the methods. The means and standard deviations of the generalisation error for each method and dataset are given in Table 1. We find that the performance of KFDA and MPKFDA are very similar, and both are competitive with the SVM. This is demonstrated by the values for the mean over the datasets.

Next we present results from the NIPS 2003 challenge datasets [Guyon et al., 2004] ARCENE, DEXTER and DOROTHEA³. These datasets were chosen as we believe that the main advantage of MPKFDA will be shown when the data lives in high dimensions. We compare the performance of MPKFDA with standard KFDA and SVM, again using an RBF kernel for each of the classifiers. We used 5-fold cross validation on the training set to select the optimal parameters for each algorithm as before, and then tested on the validation set. For each dataset we show the number of features, and the number of examples in the training and validation sets, and the generalisation error of each classifier on the validation set. All problems are two-class classification problems. As can be seen from Table 2, MPKFDA outperforms both KFDA and SVM on these high dimensional datasets, whilst giving very sparse solutions.

6 CONCLUSIONS

In this paper we derived a novel sparse version of Kernel Fisher Discriminant Analysis (KFDA) using an approach based on Matching Pursuit (MP). We provided generalisation error bounds analogous to that used in the Robust Minimax algorithm [Lanckriet et al., 2003], together with a sample compression bounding technique. As it stands the bound is too loose to perform model selection, but will anticipate that further analysis may enable the bound to drive the algorithm. We presented experimental results on real world datasets, which showed that MPKFDA is competitive with both KFDA and SVM, and additional experiments that showed that MPKFDA performs extremely well in high dimensional settings. In terms of computational complexity the demands of MPKFDA during training are higher, but during the evaluation

on test points only k kernel evaluations are required compared to m needed for KFDA.

The Fisher discriminant is the Bayes optimal classifier for two normal distributions with equal covariance Kim et al. [2006]. This probabilistic interpretation can be extended to the Matching Pursuit setting, where our input space is now the compressed space induced by the projections. [Kim et al., 2006] showed that Robust Fisher Discriminant Analysis was able to mitigate against the sensitivity to problem data in FDA by explicitly incorporating a model of data uncertainty into the classification problem and optimising for the worst-case scenario under this model. We believe this method could also be applied to MPKFDA.

We believe this general approach of using Matching Pursuit can be applied to other learning algorithms, resulting in sparse greedy forms of these algorithms. It would be conceivable to apply the method to Logistic Regression. FDA has an advantage over Logistic Regression that it has a probabilistic model of both positive and negative data, which may prove useful in further analysis. This also suggests a very natural way to extend this work to multi-class classification.

Speeding up of the algorithm is also considered an important future research direction and not something we paid particular attention to throughout this work. The quotient that we maximise in Line 4 of the MPKFDA Algorithm serves as a reference point for future studies and currently requires m^3 computations at each step. Relieving this issue could speed up the algorithm considerably and allow its application to a much larger class of dataset.

Traditional OMP and KMP with pre-fitting work by incrementing each element of the weight vector after each choice of basis vector. However, we constructed the full set k of bases before computing the final weight vector. Although this did not prove detrimental we still feel that addressing this issue would make us more computationally favourable and allow the algorithm to be fully incremental in its greedy strategy.

Another issue of speeding up the algorithm may be to consider approximating ρ . A fast and quality approximation would yield much faster convergence rates and also still be amenable to the bounds we presented within this paper.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under *grant agreement* N° 216529, Personal Information Navigator Adapting Through Viewing, PinView and the EP-

³The train and validation sets and associated labels are available for download from:

<http://www.nipsfsc.ecs.soton.ac.uk/datasets/>

Table 1: Generalisation error estimates and standard deviations for 13 benchmark datasets.

	Dim	Train	Test	KFDA		MPKFDA			SVM	
				Error	s.d.	Error	s.d.	k	Error	s.d.
Banana	2	400	4900	0.1069	0.0047	0.1101	0.0071	31	0.1068	0.0047
Breast Cancer	9	200	77	0.2886	0.0468	0.3174	0.0447	19	0.2603	0.0473
Diabetes	8	468	300	0.2596	0.0203	0.2543	0.0189	18	0.2332	0.0175
Flare Solar	9	666	400	0.3500	0.0168	0.3457	0.0220	19	0.3239	0.0179
German	20	700	300	0.2672	0.0248	0.2808	0.0205	27	0.2345	0.0215
Heart	13	170	100	0.2125	0.0327	0.1599	0.0312	13	0.1543	0.0326
Image	18	1300	1010	0.0092	0.0187	0.0136	0.0278	39	0.0061	0.0124
Ringnorm	20	400	7000	0.0685	0.0108	0.0573	0.0302	15	0.0164	0.0012
Splice	60	1000	2175	0.0397	0.0801	0.0314	0.0633	37	0.0223	0.0450
Thyroid	5	140	75	0.0392	0.0208	0.0699	0.0310	29	0.0520	0.0208
Titanic	3	150	2051	0.2259	0.0247	0.2468	0.0528	70	0.2256	0.0110
Twonorm	20	400	7000	0.0253	0.0022	0.0253	0.0016	14	0.0280	0.0024
Waveform	21	400	4600	0.1228	0.0053	0.1027	0.0046	13	0.1031	0.0047
Mean				0.1550	0.0237	0.1550	0.0274	26.5	0.1359	0.0184

Table 2: Generalisation error estimates for 3 high dimensional datasets.

	Dim	Train	Test	KFDA	MPKFDA		SVM
				Error	Error	k	Error
Arcene	10000	100	100	0.2000	0.1800	40	0.2600
Dexter	20000	300	300	0.1133	0.0800	40	0.0733
Dorothea	100000	800	350	0.0971	0.0571	11	0.0686
Mean				0.1368	0.1057	30.3	0.1340

SRC grant agreement N° EP-D063612-1, Learning the Structure of Music.

References

- Francis R. Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 33–40, New York, NY, USA, 2005. ACM.
- Peter L. Bartlett and Ambuj Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8: 775–790, 2007.
- Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
- Murat Dunder, Glenn Fung, Jinbo Bi, Sandilya Sathyakama, and Bharat Rao. Sparse fisher discriminant analysis for computer aided detection. In *Proceedings of the SIAM International Conference on Data Mining*, 2005.
- Yuanjian Feng and Pengfei Shi. Face detection based on kernel fisher discriminant analysis. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.
- Roland .A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
- Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.
- Isabelle Guyon, Asa Ben Hur, Steve Gunn, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT Press, 2004.
- Zakria Hussain and John Shawe-Taylor. Theory of matching pursuit. *Neural Information Processing Systems*, 2008.
- Seung-Jean Kim, Alessandro Magnani, and Stephen P. Boyd. Robust fisher discriminant analysis. In *In Advances in Neural Information Processing Systems*, pages 659–666. MIT Press, 2006.
- Brian Kulis, Máttyás Sustik, and Inderjit Dhillon. Learning low-rank kernel matrices. In *ICML '06: Proceedings of the 23rd international conference on*

- Machine learning*, pages 505–512, New York, NY, USA, 2006. ACM.
- Gert R.G. Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. A robust minimax approach to classification. *J. Mach. Learn. Res.*, 3:555–582, 2003. ISSN 1533-7928.
- Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.
- Stéphane Mallat and Zhifeng Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
- S. Mika, A. J. Smola, and B. Schölkopf. An improved training algorithm for kernel fisher discriminants. In *AISTATS*, pages 98–104, 2001.
- Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert Müller. Fisher discriminant analysis with kernels. In E. Wilson Y. H. Hu, J. Larsen and S. Douglas, editors, *Proc. NNSP'99*, pages 41–48. IEEE, 1999.
- Yagyensh Chandra Pati, Ramin Rezaifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–45, 1993.
- John Shawe-Taylor and Nello Cristianini. Estimating the moments of a random vector. In *Proceedings of GRETSI 2003 Conference*, volume 1, page 4752, 2003.
- John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.
- Alex J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, San Francisco, CA, 2000.
- Lelsie G. Valiant. A theory of the learnable. *Communications of the Association of Computing Machinery*, 27(11):1134–1142, November 1984.
- Pascal Vincent and Yoshua Bengio. Kernel matching pursuit. *Machine Learning*, 48(1-3):165–187, 2002.
- Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pages 682–688. MIT Press, 2001.
- Hong-Jie Xing, Yu-Jiu Yang, Yong Wang, and Bao-Gang Hu. Sparse kernel fisher discriminant analysis. *Springer Lecture Notes in Computer Science*, 3496: 824–830, 2005.