
Monograph:
Lee, K.L. and Billings, S. (2000) Nonlinear Fisher Discriminant Analysis Using a Minimum Squared Error Cost Function and the Orthogonal Least Squares Algorithm. Research Report. ACSE Research Report 781 . Department of Automatic Control and Systems Engineering

---

Nonlinear Fisher Discriminant Analysis using a Minimum Squared Error Cost Function and the Orthogonal Least Squares Algorithm

K.L. Lee S.A. Billings
Department of Automatic Control and Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
United Kingdom

Research Report No. 781

October 2000

University of Sheffield

Nonlinear Fisher Discriminant Analysis using a Minimum Squared Error Cost Function and the Orthogonal Least Squares Algorithm

K.L. Lee
Email: cop98kll@sheffield.ac.uk

S.A. Billings
Email: s.billings@sheffield.ac.uk

Department of Automatic Control
and Systems Engineering
University of Sheffield
Sheffield S1 3JD, UK

Abstract: The nonlinear discriminant function obtained using a minimum squared error cost function can be shown to be directly related to the nonlinear Fisher discriminant. With the squared error cost function, the orthogonal least squares algorithm can be used to find a parsimonious description of the nonlinear discriminant function. Two simple classification techniques will be introduced and tested on a number of real and artificial data sets. The results show that the new classification technique can often perform favourably compared with other state of the art classification techniques.

Keywords: Pattern classification, Fisher discriminant, Orthogonal least squares algorithm, Squared error cost function, nonlinear kernel functions, parsimonious, Kernel fisher discriminant, Regularisation.

Nomenclature

$\mathbf{x}_i$	Input sample i.
$\mathbf{C}_i$	Class i.
n	Total number of input samples.
n_i	Number of input samples for class i.
$\mathbf{w}$	(Non)linear discriminant function.
S_B	Between-class scatter matrix.
S_W	Within-class scatter matrix.
$\mathbf{m}_i$	Sample mean of class i.
$\Phi(\cdot)$	Nonlinear function.
F	High dimensional feature space.
S_B^Φ	Between-class scatter matrix in feature space.
S_W^Φ	Within-class scatter matrix in feature space.
$\mathbf{m}_i^\Phi$	Sample mean of class i in feature space.

$k(\cdot,\cdot)$	Mercer kernel.
d	Degree of nonlinearity for polynomial kernel.
σ	Width of the gaussian kernel.
α	Expansion coefficients of the discriminant function.
W	Classification function.
Ξ	Error vector.
Y	Target vector.
y_i	Target value for class i.
u_i	Column vector of n_i ones.
w_o	Threshold.
P	Kernel matrix.
T	$n \times n$ Orthogonal matrix.
A	Upper unit triangular matrix.
Θ	Weights of the classification function.

1. Introduction

Fisher linear discriminant analysis is widely known and used in practice. However, linear discriminant analysis is certainly not complex enough for most real world data, hence it is important to develop nonlinear discriminant analysis methods. Recently, kernel based algorithms (Mika et al. 1999) have been proposed to define a nonlinear generalisation of the Fisher linear discriminant algorithm for pattern classification problems. By reformulating the problem into dot product form in a higher dimensional space and using kernels which satisfy the Mercer condition (Vapnik 1995), a closed form solution of the nonlinear discriminant function has been obtained. More importantly, very promising results were reported with the use of the nonlinear Kernel Fisher Discriminant (KFD) algorithm (Mika et al. 1999) when compared with the other state of the art classification techniques. However one particular drawback with the method is that the complexity of the nonlinear Fisher discriminant function scales with the number of training data. This implies that the testing time is going to be very slow for problems with a large number of training data and the computational cost for storing a large data set is also going to be high.

By using a minimum squared error cost function for pattern classification problems, the resulting linear discriminant is directly related to the Fisher linear discriminant (Duda and Hart 1973). With the use of a nonlinear mapping of the input samples, the above relationship can easily be extended to the nonlinear

case. Therefore the nonlinear discriminant function obtained by minimising the squared error cost function is directly related to the nonlinear Fisher discriminant. Hence with a minimum squared error cost function, the well-known Orthogonal Least Squares (OLS) algorithm (Chen et al. 1989 and Chen et al. 1991) can be used to find a parsimonious description for the nonlinear discriminant function. In addition the resulting algorithm is simpler to compute than the Kernel Fisher Discriminant (KFD).

The paper is organised as follows. In section 2 a brief review of the linear and Kernel Fisher Discriminant (KFD) algorithms is given. The direct relationship between the nonlinear discriminant obtained by minimising a squared error cost function and the nonlinear Fisher discriminant is shown in section 3. A brief review of the orthogonal least squares algorithm is also given in Section 3. Two simple classification methods will be introduced using the orthogonal least squares algorithm and these two methods are tested on an extensive number of experiments and the results are presented in section 4. Brief conclusions are provided in Section 5.

2. Linear and Kernel Fisher Discriminants

2.1 Linear Fisher Discriminant Analysis

Given a set of n d-dimensional samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \mathbf{x}_i \in \Re^d\}$ with n_1 samples in Class 1 denoted $\mathbf{C}_1$ and n_2 samples in class 2 denoted $\mathbf{C}_2$ then the Fisher linear discriminant (Duda and Hart 1973 and Ripley 1996) is given by the vector $\mathbf{w}$, $(\mathbf{w} \in \Re^d)$ which maximises the following function

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

where $S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ (2)

and $S_W = \sum_{i=1,2} \sum_{\mathbf{x} \in \mathbf{C}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$ (3)

S_B is known as the between-class scatter matrix, S_W is called the within-class scatter matrix and $\mathbf{m}_i$ is the sample mean of the respective classes defined as

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathbf{C}_i} \mathbf{x} \tag{4}$$

The reasoning behind making $J(\mathbf{w})$ as large as possible is to look for a direction $\mathbf{w}$ which maximises the difference between the two projected means in S_B while minimising the variance of the individual classes S_W. Hence samples belonging to two different classes are well separated by projection onto this optimal direction. This effect is illustrated in Figure 1. Furthermore, by assuming normal distributions

and equal covariance for the two different classes, the resulting linear discriminant function is in the same direction as the Bayes optimal classifier.

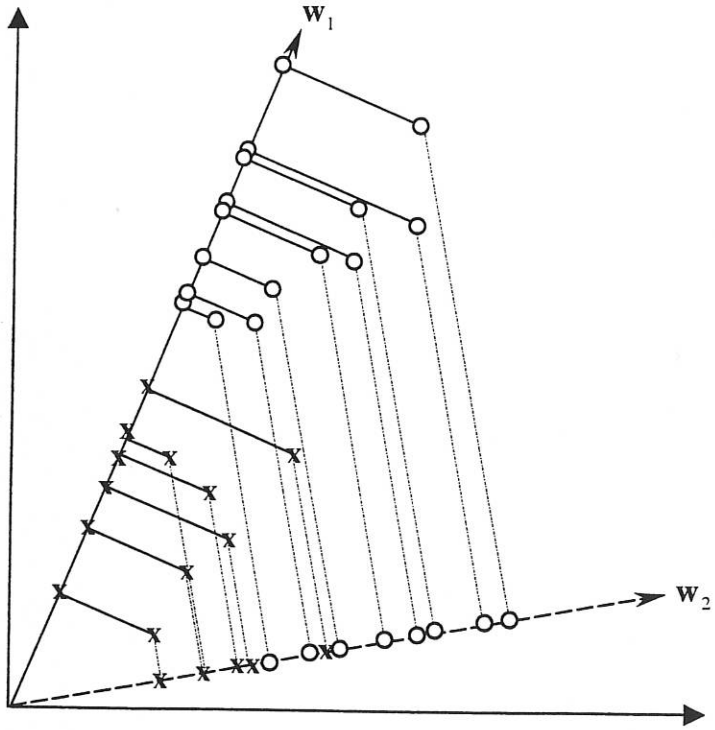


Figure 1. Projection of 2-dimensional samples onto a line. In direction $\mathbf{w}_1$ the projected samples are well separated whereas in direction $\mathbf{w}_2$ the projected samples are mixed.

2.2 Nonlinear Fisher Discriminant Analysis

However, the usefulness of linear Fisher discriminant analysis is somewhat limited in real world problems and a nonlinear Fisher discriminant analysis would be highly desirable. A simple method of obtaining the nonlinear discriminant is to first map the samples into some high dimensional feature space F using a nonlinear function $\Phi(\cdot)$, then linear discriminant analysis can be performed in this feature space. The resulting discriminant function will be linear in the feature space F but nonlinear in the input space. Hence, the nonlinear discriminant function $\mathbf{w} \in F$ can be obtained by maximising

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B^\Phi \mathbf{w}}{\mathbf{w}^T S_W^\Phi \mathbf{w}} \tag{5}$$

where $S_B^\Phi = \left(\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi\right)\left(\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi\right)^T$ (6)

and $S_W^\Phi = \sum_{i=1,2}\sum_{\mathbf{x}\in\mathbf{C}_i}\left(\Phi(\mathbf{x}) - \mathbf{m}_i^\Phi\right)\left(\Phi(\mathbf{x}) - \mathbf{m}_i^\Phi\right)^T$ (7)

$$\mathbf{m}_i^{\Phi} = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \Phi(\mathbf{x}) \tag{8}$$

Eqn (5) is a well-known expression called the generalised Rayleigh quotient and a vector $\mathbf{w}$ which maximises $J(\mathbf{w})$ must satisfy

$$\gamma S_W^{\Phi} \mathbf{w} = S_B^{\Phi} \mathbf{w} \tag{9}$$

where γ is a constant. Assuming that S_W^{Φ} is nonsingular

$$\gamma \mathbf{w} = \left(S_W^{\Phi}\right)^{-1} S_B^{\Phi} \mathbf{w} \tag{10}$$

then eqn (10) can be solved by finding the eigenvectors of $(S_W^{\Phi})^{-1} S_B^{\Phi}$. But this may not be necessary since

$$\begin{aligned} S_B^{\Phi} \mathbf{w} &= \left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right)\left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right)^T \mathbf{w} \\ &= \beta\left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right), \ \beta \text{ is a constant.} \end{aligned} \tag{11}$$

Hence, the discriminant function $\mathbf{w}$ becomes

$$\mathbf{w} = \left(S_W^{\Phi}\right)^{-1} \left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right) \tag{12}$$

Only the direction is important, so the constant is dropped in eqn (12). Unfortunately, the mapping function $\Phi(\cdot)$ may not be known explicitly and if the dimension of the feature space F is very high or infinite, eqn (12) may be difficult to use to solve for the discriminant function. To get around this difficulty, the problem is reformulated to involve only the dot product of the training samples in the feature space, that is $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Furthermore, using Mercer kernels (Vapnik 1995) which satisfy

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \tag{13}$$

The problem can be solved without ever mapping explicitly to the feature space F.

The polynomial kernel $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$, where d is the degree of nonlinearity, and the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\| / \sigma^2\right)$, where σ is the width are some typical kernels which satisfy the Mercer condition.

Following (Mika et al. 1999), since the nonlinear discriminant function $\mathbf{w} \in F$, this must lie in the span of all training samples in F according to the theory of reproducing kernels.

Hence $\mathbf{w}$ can be expressed as

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i) \tag{14}$$

After some manipulation of equations, eqn (5) becomes

maximising $J(\alpha) = \dfrac{\alpha^T M \alpha}{\alpha^T N \alpha}$ (15)

where $M = (M_1 - M_2)(M_1 - M_2)^T$ with $M_i = \dfrac{1}{n_i} \sum_{j=1}^{n_i} \sum_{\mathbf{x} \in \mathbf{C}_i} k(\mathbf{x}_j, \mathbf{x})$

and $N = \sum_{i=1,2} K_i (I - 1_{n_i}) K_i^T$

K_i is a $n \times n_i$ matrix with elements $(K_i)_{ab} = k(\mathbf{x}_a, \mathbf{x}_b)$, $\mathbf{x}_a, \mathbf{x}_b \in \mathbf{C}_i$. I is the identity matrix and 1_{n_i} is the matrix with all elements set to $1/n_i$.

Eqn (15) can be solved by finding the leading eigenvector of $N^{-1}M$ and the projection of a new sample $\mathbf{x}$ onto $\mathbf{w}$ is given by

$$\Phi(\mathbf{x}) \cdot \mathbf{w} = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \quad (16)$$

However, the discriminant function obtained is only an optimal direction which separates the samples from the two different classes. To solve the classification problem, an optimal threshold needs to be determined. In (Mika et al. 1999) a linear Support Vector Machine (SVM) (Vapnik 1995 and Burges 1998) was employed to estimate the optimal threshold.

Although very promising results were obtained with the use of the Kernel Fisher Discriminant (KFD) algorithm, there are a few drawbacks with this method. For problems with a large number of training samples, it is very computationally intensive to find the leading eigenvectors of $N^{-1}M$. Also, from eqn (16), it can be seen that the complexity of the discriminant function scales with the number of training samples. This is highly undesirable and this approach will be slow in the testing phase. Furthermore, linear Support Vector Machines (SVM) or other classification techniques have to be used to estimate the optimal threshold, this introduces an extra step into the method. The use of SVM also means another extra parameter has to be controlled, the regularisation constant in SVM. Note that regularisation has already been used in the KFD algorithm by replacing N with N_μ

$$N_\mu = N + \mu I \quad (17)$$

where I is again the identity matrix and μ is a positive constant.

The above mentioned problems can be avoided by employing a minimum squared error cost function and using the OLS algorithm (Chen et al. 1991). The new algorithm that is produced using this approach is discussed in the next section.

3. Minimum Squared Error and the Orthogonal Least Squares Algorithm

3.1 Minimum squared error cost function

Given a set of n d-dimensional samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \mathbf{x}_i \in \Re^d\}$, with no loss of generality, assume that the first n_1 samples are in class 1 denoted with values y_1 and the next n_2 samples are in class 2 denoted with values y_2. A nonlinear map $\Phi(\cdot)$ is applied to the samples to give

$$\begin{bmatrix} 1 & \Phi(\mathbf{x}_1) \\ \vdots & \vdots \\ 1 & \Phi(\mathbf{x}_{n_1}) \\ 1 & \Phi(\mathbf{x}_{n_1+1}) \\ \vdots & \vdots \\ 1 & \Phi(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_1 \\ y_2 \\ \vdots \\ y_2 \end{bmatrix} \tag{18}$$

where w_0 is the threshold, $\mathbf{w}$ is the discriminant function and $e_i, i = 1,2...n$ are the errors.

In matrix form: $XW + \Xi = Y$ (19)

The task is to find the classification function W which minimises the square of the error

$$J(W) = \|\Xi\|^2 = \|Y - XW\|^2 \tag{20}$$

A well-known solution is

$$X^T XW = X^T Y \tag{21}$$

$$W = (X^T X)^{-1} X^T Y \tag{22}$$

3.2 Relation to nonlinear Fisher discriminant analysis

The linear discriminant function obtained using a minimum squared error cost function has been shown to be directly related to the linear Fisher discriminant (Duda and Hart 1973). For the nonlinear case, the same result should apply since a linear discriminant will still be performed in the feature space and the nonlinearity will simply be derived from the nonlinear mapping $\Phi(\cdot)$ of the input data. Following the linear case by setting $y_1 = \frac{n}{n_1}$ and $y_2 = -\frac{n}{n_2}$, it can be shown that the minimum squared error discriminant function $\mathbf{w}$ is directly related to the nonlinear Fisher discriminant.

Let $\mathbf{u}_i$ be a column vector of n_i ones and $\Phi(X_1) = [\Phi(\mathbf{x}_1) \cdots \Phi(\mathbf{x}_{n_1})]^T$ and $\Phi(X_2) = [\Phi(\mathbf{x}_{n_1+1}) \cdots \Phi(\mathbf{x}_n)]^T$, so that the matrix X becomes

$$X = \begin{bmatrix} \mathbf{u}_1 & \Phi(X_1) \\ \mathbf{u}_2 & \Phi(X_2) \end{bmatrix} \tag{23}$$

Hence eqn (21) takes the form

$$\begin{bmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T \\ \Phi(\mathrm{X}_1)^T & \Phi(\mathrm{X}_2)^T \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 & \Phi(\mathrm{X}_1) \\ \mathbf{u}_2 & \Phi(\mathrm{X}_2) \end{bmatrix} \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T \\ \Phi(\mathrm{X}_1)^T & \Phi(\mathrm{X}_2)^T \end{bmatrix} \begin{bmatrix} \frac{n}{n_1}\mathbf{u}_1 \\ -\frac{n}{n_2}\mathbf{u}_2 \end{bmatrix} \tag{24}$$

Since $\mathbf{m}_i^{\Phi}$ is defined as

$$\mathbf{m}_i^{\Phi} = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathbf{C}_i} \Phi(\mathbf{x}) \tag{25}$$

and $$S_W^{\Phi} = \sum_{i=1,2} \sum_{\mathbf{x} \in \mathbf{C}_i} \left(\Phi(\mathbf{x}) - \mathbf{m}_i^{\Phi}\right)\left(\Phi(\mathbf{x}) - \mathbf{m}_i^{\Phi}\right)^T \tag{26}$$

Then eqn (24) can be simplified further to give

$$\begin{bmatrix} n & \left(n_1\mathbf{m}_1^{\Phi} + n_2\mathbf{m}_2^{\Phi}\right)^T \\ \left(n_1\mathbf{m}_1^{\Phi} + n_2\mathbf{m}_2^{\Phi}\right) & S_W^{\Phi} + n_1\mathbf{m}_1^{\Phi}\left(\mathbf{m}_1^{\Phi}\right)^T + n_2\mathbf{m}_2^{\Phi}\left(\mathbf{m}_2^{\Phi}\right)^T \end{bmatrix} \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} 0 \\ n\left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right) \end{bmatrix} \tag{27}$$

Solving eqn (27) gives

$$w_0 = -\left(\mathbf{m}^{\Phi}\right)^T \mathbf{w} \tag{28}$$

$$\mathbf{w} = \eta\left(S_W^{\Phi}\right)^{-1}\left(\mathbf{m}_1^{\Phi} - \mathbf{m}_2^{\Phi}\right) \tag{29}$$

where η is a constant and $\mathbf{m}^{\Phi}$ is the total sample mean in the feature space and is defined as

$$\mathbf{m}^{\Phi} = \frac{n_1\mathbf{m}_1^{\Phi} + n_2\mathbf{m}_2^{\Phi}}{n} \tag{30}$$

Clearly both eqn (12) and eqn (29) are similar except for an unimportant constant. This shows that the nonlinear discriminant function obtained by using a minimum squared error cost function with $y_1 = \frac{n}{n_1}$ and $y_2 = -\frac{n}{n_2}$ is identical to the nonlinear Fisher discriminant case.

However, eqn (22) cannot be used to solve for the classification function W since the matrix X may not be known explicitly. Therefore, a similar trick of transforming the problem into dot product form is required in this case as well. Substituting eqn (14) into eqn (18) and replacing $\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$ with $k(\mathbf{x}_i, \mathbf{x}_j)$ gives

$$\begin{bmatrix} 1 & k(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & k(\mathbf{x}_n,\mathbf{x}_1) & \cdots & k(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} w_o \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} \frac{n}{n_1}\mathbf{u}_1 \\ -\frac{n}{n_2}\mathbf{u}_2 \end{bmatrix} \tag{31}$$

in matrix form

$$P\Theta + \Xi = \mathrm{Y} \tag{32}$$

In this case the matrix P is known explicitly and the problem is to find Θ which minimises the squared error. The solution for Θ can be obtained as

$$\Theta = \left(P^T P\right)^{-1} P^T \mathrm{Y} \tag{33}$$

Typically the matrix $\left(P^T P\right)$ is highly ill conditioned and one solution is to employ regularisation, where the matrix $\left(P^T P\right)$ is replaced by $\left(P^T P + \lambda I\right)$ where λ is a positive constant. Therefore the threshold and the discriminant function can be obtained by solving

$$\Theta = \left(P^T P + \lambda I\right)^{-1} P^T \mathrm{Y} \tag{34}$$

Knowing the values of Θ and letting c be the mid-value of $\frac{n}{n_1}$ and $-\frac{n}{n_2}, c = \frac{1}{2}\left(\frac{n}{n_1} - \frac{n}{n_2}\right)$, then the following decision rule will be obtained. For a new sample $\mathbf{x}$, decide class 1 if $w_0 + \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) > c$, otherwise decide class 2. Hence there is no requirement to use an additional classification technique to determine the threshold.

This formulation avoids the two problems in KFD, firstly in finding the leading eigenvectors of $N^{-1}M$ and secondly in using an additional classification technique to estimate the threshold. However this approach is not pursued in this study since the solution obtained is not sparse, in the sense that all the training samples are used to construct the classification function.

The answer in obtaining a sparse solution can be provided by using the well-known Orthogonal Least Squares (OLS) algorithm. This new method of finding the classification function by using a Nonlinear Fisher Discriminant function combined with the OLS algorithm will be called the NFD-OLS algorithm.

3.3 The Orthogonal Least Squares Algorithm

A brief review of the Orthogonal Least Squares algorithm (Chen et al. 1989 and Chen et al. 1991) is given below.

An orthogonal decomposition of P is given as

$$P = TA \tag{35}$$

where A is an $n \times n$ upper unit triangular matrix.

$$A = \begin{bmatrix} 1 & \tau_{12} & \cdots & \tau_{1n} \\ & 1 & \cdots & \tau_{2n} \\ & & \ddots & \vdots \\ 0 & & & 1 \end{bmatrix} \tag{36}$$

and T is a $n \times n$ matrix with orthogonal columns satisfying

$$T^T T = diag\{\kappa_1, \kappa_2, ... \kappa_n\}, \ \kappa_i = t_i^T t_i \tag{37}$$

Rearranging eqn (32) yields

$$\mathrm{Y} = (PA^{-1})(A\Theta) + \Xi = Tg + \Xi \tag{38}$$

With a squared error cost function, the values of g can be obtained easily as

$$g_i = \frac{t_i^T \mathrm{Y}}{t_i^T t_i} \tag{39}$$

and $A\Theta = g$ (40)

Knowing A and g, Θ can be determined through back substitution.

To obtain the Error Reduction Ratio (ERR), consider the squared error cost function

$$J = \Xi^T \Xi = (\mathrm{Y} - Tg)^T (\mathrm{Y} - Tg) \tag{41}$$

Using eqn (39), the above eqn (41) can be simplified as

$$J = \mathrm{Y}^T \mathrm{Y} - g^T T^T T g \tag{42}$$

Hence Error Reduction Ratio (ERR) due to P_i may be expressed as

$$ERR_i = \frac{g_i^2 t_i^T t_i}{\mathrm{Y}^T \mathrm{Y}} \tag{43}$$

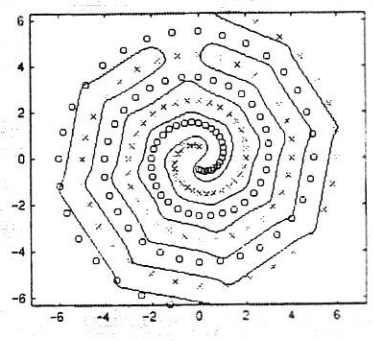
Using this ERR ratio, significant terms (columns of the matrix P) can be selected in a forward-selection procedure (Chen et al. 1989, Chen et al. 1991). At the i iteration, the term which gives the largest value of ERR_i is selected and added to the previously selected $(i-1)$ terms. The selection procedure can be terminated using cross-validation. Empirical results in time series analysis have shown that parsimonious models can be obtained using the orthogonal least squares algorithm.

3.4 A side note on other values of Y

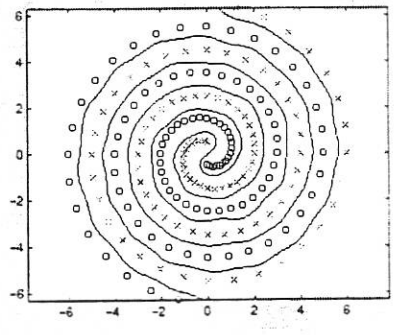
Instead of putting $y_1 = \frac{n}{n_1}$ and $y_2 = -\frac{n}{n_2}$ to differentiate the two classes, y_1 and y_2 can simply be chosen as +1 and –1 respectively. The task is to find the underlying relationship, which maps the input data to the correct classes. The following decision rule will be obtained. For a new sample $\mathbf{x}$, decide class 1 if $w_0 + \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) > 0$, otherwise decide class 2. This is called the Simple OLS method.

4. Experiments

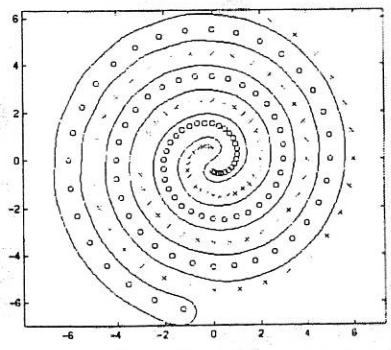
The two spirals problem (Lang and Witbrock 1988) is chosen as the first experiment to test the new Nonlinear Fisher Discriminant with OLS algorithm (NFD-OLS). The two spiral problem is a well-known benchmark problem and the two dimensional nature of the problem allows easy visualisation of the classification function learned by the algorithm. The kernel chosen was Gaussian with width equal to 1. There were 194 training samples and the result is displayed in Figure 2. Clearly, the number of nodes used to construct the classification function will play an important role. It can be seen that with 120 nodes, the classification function produces large margins between the decision function and the training samples. This implies a good generalisation property. Note that in this two spirals example, the KFD or SVM algorithms used all the 194 training samples in constructing the classification function.



(a) 50 nodes



(b) 90 nodes



(c) 120 nodes

Figure 2 .Two spirals classification problem with 194 training samples. The kernel used is the Gaussian kernel with a width equal to 1. In (a) the number of nodes chosen by OLS to construct the classification function was 50, whereas in (b) and (c) there were 90 and 120 nodes respectively.

In order to test the usefulness of the Simple OLS and NFD-OLS algorithms thoroughly, an extensive number of experiments were conducted on 13 artificial and real world datasets[1]. The datasets used in (Mika et al. 1999 and Rätsch et al. 1998) were chosen. These datasets can be downloaded from http://www.first.gmd.de/~raetsch/. For each of the 13 cases (except Splice and Image with 20 partitions), a further 100 random partitions were generated. In all cases, Gaussian kernels were used and the width and optimal number of nodes chosen by OLS were estimated from the first five realisations of the training and test samples using 5-fold cross validation. The width and the number of nodes, which gave the minimum total classification error on the first five test sets were chosen to construct the optimal classification function. This optimal classification function was then used on all the 100 partitions and the mean classification errors of the test sets for the 13 cases are shown in Table 1.

From Table 1, the results obtained using Simple OLS and Nonlinear Fisher Discriminant with OLS (NFD-OLS) are quite favourable compared to the other existing classification methods. The Nonlinear Fisher dicriminant with OLS algorithm is slightly superior than the other methods since this approach produced the best results four times (highest) out of the 13 cases considered. Note that the number of nodes selected by Simple OLS and NFD-OLS to construct the classification functions for the 13 cases were in most cases less than 10% of the training data. The exact numbers of nodes used are shown in Table 2. Hence a very economical solution was obtained using the OLS algorithm in pattern classification problems and this produced short testing times and a saving in storage costs.

[1] The Breast Cancer dataset was provided by the University Medical Center, Inst. of Oncology, Ljubljana. Thanks to M. Zwitter and M. Soklic for the dataset.

Data Set	RBF	AB	AB_R	SVM	KFD	Simple OLS	NFD-OLS
Banana	*10.8±0.6*	12.3±0.7	10.9±0.4	11.5±0.7	*10.8±0.5*	*10.8±0.4*	**10.7±0.5**
B. Cancer	27.6±4.7	30.4±4.7	26.5±4.5	26.0±4.7	*25.8±4.6*	**25.3±4.1**	*25.8±4.8*
Diabetes	24.3±1.9	26.5±2.3	23.8±1.8	23.5±1.7	*23.2±1.6*	23.3±1.8	**23.1±1.8**
German	24.7±2.4	27.5±2.5	24.3±2.1	**23.6±2.1**	*23.7±2.2*	24.0±2.1	24.0±2.3
Heart	17.6±3.3	20.3±3.4	16.5±3.5	*16.0±3.3*	16.1±3.4	16.5±3.1	**15.8±3.4**
Image	3.3±0.6	**2.7±0.7**	**2.7±0.6**	3.0±0.6	4.8±0.6	*2.8±0.6*	2.9±0.5
Ringnorm	1.7±0.2	1.9±0.3	*1.6±0.1*	1.7±0.1	**1.5±0.1**	*1.6±0.1*	*1.6±0.1*
F. Sonar	34.4±2.0	35.7±1.8	34.2±2.2	**32.4±1.8**	*33.2±1.7*	33.5±1.6	33.6±1.6
Splice	*10.0±1.0*	10.1±0.5	**9.5±0.7**	10.9±0.7	10.5±0.6	11.8±0.6	11.7±0.6
Thyroid	4.5±2.1	*4.4±2.2*	4.6±2.2	4.8±2.2	**4.2±2.1**	4.5±2.4	4.6±2.4
Titanic	23.3±1.3	*22.6±1.2*	*22.6±1.2*	**22.4±1.0**	23.2±2.0	*22.6±1.1*	**22.4±1.0**
Twonorm	2.9±0.3	3.0±0.3	*2.7±0.2*	3.0±0.2	**2.6±0.2**	*2.7±0.2*	*2.7±0.2*
Waveform	10.7±1.1	10.8±0.6	**9.8±0.8**	*9.9±0.4*	*9.9±0.4*	10.0±0.4	10.0±0.4

Table 1. Comparison on the mean test set classification errors between the Simple OLS, Nonlinear Fisher Discriminant with OLS (NFD-OLS), Kernel Fisher Discriminant (KFD), a single RBF classifier (RBF), AdaBoost (AB), regularised AdaBoost (AB_R) and Support Vector Machine (SVM) algorithms. Details of the other methods can be found in (Rätsch et al. 1998). The best method is highlighted in bold and the second best in italic. The mean test set classification errors with standard deviation are also shown.

Data Set	No of training samples	No of nodes chosen by Simple OLS	No of nodes chosen by NFD-OLS
Banana	400	23	29
B. Cancer	200	7	6
Diabetes	468	5	10
German	700	8	8
Heart	170	9	3
Image	1300	280	200
Ringnorm	400	7	9
F. Sonar	666	9	10
Splice	1000	400	330
Thyroid	140	23	23
Titanic	150	10	11
Twonorm	400	10	10
Waveform	400	27	14

Table 2. Number of nodes chosen by the OLS algorithm to construct the classification function.

5. Conclusions

The nonlinear discriminant function obtained using a minimum squared error cost function is directly related to the nonlinear Fisher discriminant. Furthermore, the threshold is obtained directly by the algorithm without the need to employ an additional classification technique. With a minimum squared error cost function, the well-known OLS algorithm can be applied to obtain an economical description of the nonlinear classification function. Two new methods the Simple OLS and the NFD-OLS were introduced to exploit these benefits.

These methods have been applied to an extensive number of real and artificial data sets. The results obtained suggest that the NFD-OLS algorithm performs at least as well and often better than other state of the art classification techniques. NFD-OLS achieved the best results 4 times out of the 13 cases considered. Furthermore, very sparse descriptions of the classification functions were obtained for most of the 13 cases studied in the experiments. This implies a very short testing time for new data. Regularisation has not been used in this study. This could easily be employed in the algorithm by using the regularised OLS formulation (Chen et al. 1996). Further improvements in the results may be obtained using regularisation. Recently another variant of the Kernel discriminant algorithm (Volker and Volker 2000) has been proposed, which is a more direct extension of the linear Fisher discriminant

and which is a multi-class classifier. However the strength of the NFD-OLS algorithm lies in the ability of the algorithm to obtain a very parsimonious structure for the discriminant function and the algorithm is very simple to use with excellent results obtained.

Acknowledgements

SAB gratefully acknowledges that part of this work was supported by EPSRC.

References

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. Vol 2 no.2, 121-167.

Chen, S. Billings, S.A. Luo, W. (1989). Orthogonal Least Squares methods and their application to non-linear system identification. International Journal of Control, Vol 50, no.5, 1873-96.

Chen, S., Chng, E.S., Alkadhimi, K. (1996). Regularized orthogonal least squares algorithm for constructing radial basis function networks International Journal of Control, vol 64, no 5, p829-837

Chen, S. Cowan, C.F.N., Grant, P.M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks, vol 2, no.2, pp302-9

Duda R.O. and Hart P.E. (1973). Pattern classification and scene analysis. New York; Chichester : Wiley-Interscience

G. Rätsch, T.Onoda, K.-R. Müller. (1998). Soft Margins for AdaBoost. Neuro COLT Technical Report TR-1998-021, Royal Holloway College. Accepted for publication in Machine Learning.

Lang, K. J. and Witbrock, M. J. (1988). Learning to tell two spirals apart. In Proceedings of the 1988 Connectionist Summer Schools. Morgan Kaufmann

Ripley, B.D. (1996) Pattern recognition and neural networks. Cambridge University Press

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller. (1999). Fisher Discriminant Analysis with Kernels. Accepted at 1999 IEEE International Workshop on Neural Networks for Signal Processing

Vapnik, V. (1995). The nature of Statistical Learning Theory, Springer Verlag.

Volker, R. and Volker, S., (2000). Nonlinear discriminat analysis using kernel functions, To appear in Advances in Neural Information Processing Systems 12. S. A. Solla, T. K. Leen, K.-R. Müller, eds., MIT Press.