# Optimally Regularised Kernel Fisher Discriminant Analysis

Kamel Saadi, Nicola L. C. Talbot and Gavin C. Cawley
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
{ks,nlct,gcc}@cmp.uea.ac.uk

## Abstract

*Mika* et al. *[3] introduce a non-linear formulation of Fisher's linear discriminant, based the now familiar "kernel trick", demonstrating state-of-the-art performance on a wide range of real-world benchmark datasets. In this paper, we show that the usual regularisation parameter can be adjusted so as to minimise the leave-one-out cross-validation error with a computational complexity of only $\mathcal{O}(\ell^2)$ operations, where $\ell$ is the number of training patterns, rather than the $\mathcal{O}(\ell^4)$ operations required for a naˇeve implementation of the leave-one-out procedure. This procedure is then used to form a component of an ef£cient heirarchical model selection strategy where the regularisation parameter is optimised within the inner loop while the kernel parameters are optimised in the outer loop.*

## 1. Introduction

Assume we are given training data $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell\} = \{\mathcal{X}_1, \mathcal{X}_2\} \subset \mathbb{R}^d$, where $\mathcal{X}_1 = \{\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_{\ell_1}^1\}$ is a set of patterns belonging to class $\mathcal{C}_1$ and similarly $\mathcal{X}_2 = \{\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_{\ell_2}^2\}$ is a set of patterns belonging to class $\mathcal{C}_2$; Fisher's linear discriminant (FLD) attempts to £nd a linear combination of input variables, $\boldsymbol{w} \cdot \boldsymbol{x}$, that maximises the average separation of the projections of points belonging to $\mathcal{C}_1$ and $\mathcal{C}_2$, whilst minimising the within class variance of the projections of those points onto the discriminant vector. The Fisher discriminant is given by the vector $\boldsymbol{w}$ maximising the Rayleigh quotient

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}, \tag{1}$$

where $\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$, is the between class scatter matrix, $\boldsymbol{m}_j$ is the mean of patterns belonging to $\mathcal{C}_j$,

$$\boldsymbol{m}_j = \frac{1}{\ell_j} \sum_{i=1}^{\ell_j} \boldsymbol{x}_i^j,$$

and $\boldsymbol{S}_W$ is the within class scatter matrix

$$\boldsymbol{S}_W = \sum_{i \in \{1,2\}} \sum_{j=1}^{\ell_i} (\boldsymbol{x}_j^i - \boldsymbol{m}_i)(\boldsymbol{x}_j^i - \boldsymbol{m}_i)^T.$$

The innovation introduced by Mika *et al.* [3] is to construct Fisher's linear discriminant in a £xed feature space $\mathcal{F}$ ($\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$) induced by a positive de£nite *Mercer* kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ de£ning the inner product $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$ (see e.g. Cristianini and Shawe-Taylor [2]). Let the kernel matrices for the entire dataset, $\boldsymbol{K}$, and for each class, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ be de£ned as follows:

$$\boldsymbol{K} = [k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell}$$

and

$$\boldsymbol{K}_i = [k_{jk}^i = \mathcal{K}(\boldsymbol{x}_j, \boldsymbol{x}_k^i)]_{j,k=1}^{j=\ell,k=\ell_i}.$$

The theory of reproducing kernels indicates that $\boldsymbol{w}$ can then be written as an expansion of the form

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i). \tag{2}$$

The objective function (1) can also be written such that the data $\boldsymbol{x} \in \mathcal{X}$ appear only within inner products, giving

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}}, \tag{3}$$

where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{\ell}$, $\boldsymbol{M} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$, $\boldsymbol{m}_i = \boldsymbol{K}_i \boldsymbol{u}_i$, $\boldsymbol{u}_i$ is a column vector containing $\ell_i$ elements with a common value of $\ell_i^{-1}$ and

$$\boldsymbol{N} = \sum_{i \in \{1,2\}} \boldsymbol{K}_i (\boldsymbol{I} - \boldsymbol{U_i}) \boldsymbol{K}_i^T,$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{U}_i$ is a matrix with all elements equal to $\ell_i^{-1}$. The coef£cients, $\boldsymbol{\alpha}$, of the expansion (2) are then given by the leading eigenvector of $\boldsymbol{N}^{-1}\boldsymbol{M}$. Note that $\boldsymbol{N}$ is likely to be singular, or at best ill-conditioned, and so a regularised solution is obtained by substituting $\boldsymbol{N}_\mu = \boldsymbol{N} + \mu\boldsymbol{I}$, where $\mu$ is a regularisation constant. To complete the kernel Fisher discriminant classi£er, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$, the bias, $b$, is given by

$$b = -\boldsymbol{\alpha}\frac{\ell_1\boldsymbol{m}_1 + \ell_2\boldsymbol{m}_2}{\ell}.$$

Xu *et al.* [7] show that the parameters of the kernel Fisher discriminant classi£er are also given by the solution of the following system of linear equations:

$$\begin{bmatrix} \boldsymbol{K}\boldsymbol{K} + \mu\boldsymbol{I} & \boldsymbol{K}\boldsymbol{1} \\ (\boldsymbol{K}\boldsymbol{1})^T & \ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{1}^T \end{bmatrix} \boldsymbol{y}, \qquad (4)$$

where $\boldsymbol{1}$ is a column vector of $\ell$ ones and $\boldsymbol{y}$ is a column vector with elements $y_i = \ell/\ell_j$ for $j = 1$ and $y_i = -\ell/\ell_j$ for $j = 2$ $\forall i : \boldsymbol{x}_i \in \mathcal{X}_j$. This illustrates the similarities between the kernel Fisher discriminant and the least-squares support vector machine (LS-SVM) [5]. The kernel Fisher discriminant (KFD) classi£er has been shown experimentally to demonstrate near state-of-the-art performance on a range of arti£cial and real world benchmark datasets [3] and so is worthy of consideration for small to medium scale applications. In this paper we present an ef£cient algorithm for approximate cross-validation of kernel Fisher discriminant models, providing a practical criterion for model selection.

## 2. Method

In this section, we describe a training algorithm for the kernel Fisher discriminant classi£er in which the system of linear equations (4) is solved in *canonical form*, allowing the regularisation parameters to be updated in only $\mathcal{O}(\ell)$ operations. An ef£cient method for approximate leave-one-out cross-validation is then presented, forming the basis of a criterion for the optimisation of the vector of regularisation parameters $\boldsymbol{\mu}$ with a complexity of only $\mathcal{O}(\ell^2)$ operations per iteration.

### 2.1. Canonical Form KFD Analysis

The system of linear equations (4) can be written more concisely in the form

$$\boldsymbol{\beta} = \left[\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{R}\right]^{-1} \boldsymbol{Z}^T\boldsymbol{y}, \qquad (5)$$

where $\boldsymbol{Z} = [\boldsymbol{K}\ \boldsymbol{1}]$ and $\boldsymbol{R}$ is a diagonal matrix with elements given by the vector of regularisation parameters $\boldsymbol{\mu}$. Let $\boldsymbol{V}$ be an orthogonal matrix, the columns of which are the eigenvectors of $\boldsymbol{Z}^T\boldsymbol{Z}$, and $\boldsymbol{\Lambda}$ be a diagonal matrix containing the corresponding eigenvalues $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_\ell \geq 0$, such that

$$\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T, \quad \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}.$$

The principal components of $\boldsymbol{Z}$ are then given by the columns of $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{V}$; note that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{\Lambda}$. The system of linear equations (5) can then be expressed in *canonical form* [6] as

$$\boldsymbol{\alpha} = \boldsymbol{C}^{-1}\boldsymbol{U}^T\boldsymbol{y} = [\boldsymbol{\Lambda} + \boldsymbol{R}]^{-1}\boldsymbol{U}^T\boldsymbol{y}, \qquad (6)$$

where $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta}$. The principal advantage of expressing the system of linear equations (5) in this form is that the matrix $\boldsymbol{C}$ is diagonal, and so can be inverted in linear time, i.e. $\mathcal{O}(\ell)$ operations.

### 2.2. Updating the Regularisation Parameter

The canonical parameters of the kernel Fisher discriminant classi£er can be written as

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\Lambda}}{[\boldsymbol{\Lambda} + \boldsymbol{R}]}\hat{\boldsymbol{\alpha}}.$$

where $\hat{\boldsymbol{\alpha}}$ are the parameters of a KFD model trained without regularisation, i.e. $\boldsymbol{\mu} = \boldsymbol{0}$. As $\boldsymbol{\Lambda}$ and $[\boldsymbol{\Lambda} + \boldsymbol{R}]$ are both diagonal matrices, we can write [6]

$$\alpha_i = \frac{\lambda_i}{\lambda_i + \mu_i}\hat{\alpha}_i, \qquad i = 0, 1, 2, \ldots, \ell,$$

It should be noted that adopting the canonical form (6), the parameters of the kernel Fisher discriminant model can be updated following a change in the vector of with a computational complexity of only $\mathcal{O}(\ell)$ operations.

### 2.3. Fast Leave-One-Out Cross-Validation

At each step of the leave-one-out cross-validation procedure, a kernel Fisher discriminant classi£er is constructed excluding a single example from the training data. The vector of canonical model parameters, $\boldsymbol{\alpha}_{(i)}$ at the $i^{\text{th}}$ step, in which pattern $i$ is excluded, is then given by the solution of a modi£ed system of linear equations,

$$\boldsymbol{\alpha}_{(i)} = \left[\boldsymbol{R} + \boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}\right]^{-1} \boldsymbol{U}_{(i)}^T\boldsymbol{y}$$

where $\boldsymbol{U}_{(i)}$ is the sub-matrix formed by omitting the $i^{\text{th}}$ row of $\boldsymbol{U}$. Note that $\boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}$ is in general no longer diagonal, and so the most computationally expensive step is again the inversion of the matrix $\boldsymbol{C}_{(i)} = \left[\boldsymbol{R} + \boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}\right]$, with a complexity of $\mathcal{O}(\ell^3)$ operations. Fortunately $\boldsymbol{C}_{(i)}$ can be written as a rank one modification of $\boldsymbol{C}$,

$$\boldsymbol{C}_{(i)} = \left[\boldsymbol{R}_{(i)} + \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{u}_i\boldsymbol{u}_i^T\right] = \left[\boldsymbol{C} - \boldsymbol{u}_i\boldsymbol{u}_i^T\right], \quad (7)$$

where $\boldsymbol{u}_i$ is the $i^{\text{th}}$ row of $\boldsymbol{U}$. This allows $\boldsymbol{C}_{(i)}^{-1}$ to be found in only $\mathcal{O}(\ell^2)$ operations, given that $\boldsymbol{C}^{-1}$ is already known, via the following matrix inversion formula : Given an invertible matrix $\boldsymbol{A}$ and column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, then assuming $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$, we have that

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

The computational complexity of the matrix inversion at each step is thus reduced from $\mathcal{O}(\ell^3)$ to $\mathcal{O}(\ell^2)$. The computational complexity of the leave-one-out cross-validation process is then only $\mathcal{O}(\ell^3)$ operations, which is the same as that of the basic training algorithm for the kernel Fisher discriminant classifier.

### 2.4. Model Selection Criterion

For model selection purposes, we are not principally concerned with the values of the model parameters themselves, but only statistics such as the leave-one-out error rate

$$E_{loo} = \frac{1}{\ell}\sum_{i=1}^{\ell}\{1 - \Psi(\text{sign}(y_i)\left\{\boldsymbol{r}_{(i)}\right\}_i + 1)\}, \quad (8)$$

where $\Psi$ is the Heaviside or unit step function,

$$\Psi(x) = \left\{\begin{array}{ll} 1 & x \geq 0 \\ 0 & x < 0 \end{array}\right. .$$

and $\left\{\boldsymbol{r}_{(i)}\right\}_i = y_i - \boldsymbol{w}_{(i)} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b_{(i)}$ is the residual error for the $i^{\text{th}}$ training pattern during the $i^{\text{th}}$ iteration of the leave-one-out cross-validation procedure. It can be shown that [1]

$$\left\{\boldsymbol{r}_{(i)}\right\}_i = \frac{1}{1 - h_{ii}} r_i.$$

where $r_i = y_i - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b$ is the residual error for the $i^{\text{th}}$ training pattern for a kernel Fisher discriminant classifier trained on the entire dataset, and $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{U}^T$ is the *hat* matrix of which $h_{ii}$ is the $i^{\text{th}}$ element of the leading diagonal [6]. In this case, $\boldsymbol{C}$ is diagonal and can be inverted in linear time, and therefore

$$h_{ii} = \sum_{j=1}^{\ell} u_{ij}^2 c_{jj}^{-1} = \sum_{j=1}^{\ell} \frac{u_{ij}^2}{(\lambda_j + \mu_j)}.$$

The leave-one-out error rate can thus be evaluated in closed form without explicit inversion of $\boldsymbol{C}_{(i)}$ $\forall i \in \{1, 2, \ldots, \ell\}$, with a computational complexity of only $\mathcal{O}(\ell^2)$ operations. To find the optimal regularisation parameters we will assume, as is normally the case, a single regularisation parameter $\mu$ and throughout the rest of this paper we choose an isotropic radial basis kernel.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right).$$

The Nelder-Mead simplex algorithm [4] is then used to search for the values of $\mu^*$ and $\sigma^*$ minimising the leave-one-out error ( 8).

## 3. Results

The proposed approximate leave-one-out cross-validation method is evaluated over a series of randomly generated synthetic datasets. The data sets vary in size between 10 and 1000 patterns. Figure 2 shows a graph of run-time as a function of the number of training patterns for direct and fast leave-one-out cross-validations and KFD training. The fast leave-one-out cross-validation is considerably faster (more than an order of magnitude) than the direct leave-one-out cross-validation and it is even faster than the training procedure itself (assuming that leave-one-out cross-validation is peformed as a bye-product of training a model on the full dataset). This suggest that it is well suited to use as a model selection method, in terms of computational complexity, for medium sized datasets.

In order to evaluate the generalisation performance of models minimising the leave-one-out error, the proposed model selection procedure was applied to 13 real-world benchmark datasets used in previous studies (e.g. [3]). Each benchmark consists of 100 random partitions of the data for form test and training sets (20 in the case of image and splice dataset). The results obtained are also compared with those of Mika et al. [3] including kernel Fisher discriminant and other state of the art classification algorithms. Table 1 shows the outcome of a comparison of the proposed optimally regularised kernel Fisher discriminant and other classsification models. The ORKFD outperforms the 10-fold cross-validation estimate of the test error rate adopted by Mika et al. (KFD) [3] on 8 of the 13 data sets(banana, diabetis, german, ringnorm, titanic, twonorm, waveform) and performs worse on the remaining five and scores the lowest error rate on 7 datasets against the other state-of-the art algorithms including KFD.

## 4. Summary

In this paper we have shown that the regularisation parameter of a kernel Fisher discriminant (KFD) classifier

| Data set | ORKFD | KFD | SVM | AdaBoostQ | AdaBoostL | RBF |
|---|---|---|---|---|---|---|
| Banana | **10.49±0.54** | 10.75±0.45 | 11.53±0.66 | 10.90±0.46 | *10.73±0.43* | 10.76±0.42 |
| Breast cancer | 26.41±4.91 | **24.77±4.63** | 26.04±4.74 | *25.91±4.61* | 26.79±6.08 | 27.64±4.71 |
| Diabetis | **23.19±1.92** | *23.21±1.63* | 23.53±1.73 | 25.39±2.20 | 24.11±1.90 | 24.29±1.88 |
| German | **23.61±1.93** | *23.71±2.20* | **23.61±2.07** | 25.25±2.14 | 24.79±2.22 | 24.71±2.38 |
| Heart | 16.19±3.37 | *16.14±3.39* | **15.95±3.26** | 17.17±3.44 | 17.49±3.53 | 17.55±3.25 |
| Image | 3.62±0.68 | 4.76±0.58 | 2.96±0.60 | **2.67±0.63** | *2.76±0.61* | 3.32±0.65 |
| Ringnorm | **1.47±0.11** | *1.49±0.12* | 1.66±0.12 | 1.86±0.22 | 2.24±0.46 | 1.70±0.21 |
| Solar flare | 34.39±1.73 | *33.16±1.72* | **32.43±1.82** | 36.22±1.80 | 34.74±2.00 | 34.37±1.95 |
| Splice | 10.92±0.73 | 10.52±0.64 | 10.88±0.66 | *10.11±0.52* | 10.22±1.59 | **9.95±0.78** |
| thyroid | 4.65±2.18 | **4.20±2.07** | 4.80±2.19 | *4.35±2.18* | 4.59±2.22 | 4.52±2.12 |
| Titanic | **22.39±1.03** | 23.25±2.05 | *22.42±1.02* | 22.71±1.05 | 23.98±4.38 | 23.26±1.34 |
| Twonorm | **2.54±0.32** | *2.61±0.15* | 2.96±0.23 | 2.97±0.26 | 3.17±0.43 | 2.85±0.28 |
| Waveform | **9.79 ±0.42** | *9.86±0.44* | 9.88±0.43 | 10.07±0.51 | 10.53±1.02 | 10.66±1.08 |

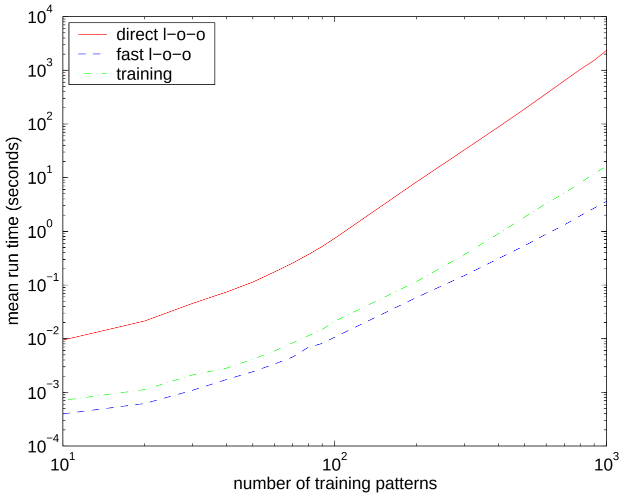

**Figure 1. Graph of run-time as a function of the number of training patterns for the KFD training algorithm, direct leave-one-out cross-validation and the fast leave-one-out cross-validation procedure.**

can be selected so as to minimise the leave-one-out cross-validation error rate, with computational complexity of only $\mathcal{O}(\ell^2)$ operations. Minimising the leave-one-out error rate then becomes an attractive model selection strategy as the scaling properties of the leave-one-out cross-validation procedure are better than those of the training procedure for the full model. The generalisation properties of models minimising the leave-one-out cross-validation are shown to be on average better than those for models minimising the more conventional 10-fold cross-validation error rate, and also superior to a range of other well-known pattern recognition algorithms.

## 5. Acknowledgements

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (BBSRC), grant number 83/EGM16128.

## References

[1] G. C. Cawley and N. L. C. Talbot. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, Nov. 2003.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, U.K., 2000.

[3] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE Press, New York, 1999.

[4] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[5] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, June 1999.

[6] S. Weisberg. *Applied linear regression*. John Wiley and Sons, New York, second edition, 1985.

[7] J. Xu, X. Zhang, and Y. Li. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. In *Proc. IJCNN*, pages 1486–1491, Washington, DC, July 2001.