

# Short Papers

## Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria

Marco Loog,

R.P.W. Duin, *Member, IEEE Computer Society,*  
and R. Haeb-Umbach, *Member, IEEE*

**Abstract**—We derive a class of computationally inexpensive linear dimension reduction criteria by introducing a weighted variant of the well-known  $K$ -class Fisher criterion associated with linear discriminant analysis (LDA). It can be seen that LDA weights contributions of individual class pairs according to the Euclidian distance of the respective class means. We generalize upon LDA by introducing a different weighting function.

**Index Terms**—Linear dimension reduction, Fisher criterion, linear discriminant analysis, Bayes error, approximate pairwise accuracy criterion.

### 1 INTRODUCTION

REDUCING the feature dimensionality of a statistical pattern classifier is a common technique to overcome estimation problems, and problems related to this. The most well-known technique for linear dimension reduction (LDR) in the  $K$ -class problem is linear discriminant analysis (LDA) (Fisher [5] introduced two-class LDA, while Rao [13] generalized LDA to multiple classes): A transformation matrix from an  $n$ -dimensional feature space to a  $d$ -dimensional space is determined such that the Fisher criterion of total scatter versus average within-class scatter is maximized [6]. Campbell has shown that the determination of the LDA transform is equivalent to finding the maximum-likelihood (ML) parameter estimates of a Gaussian model, assuming that all class discrimination information resides in a  $d$ -dimensional subspace of the original  $n$ -dimensional feature space and that the within-class covariances are equal for all classes [2].

However, for a  $K$ -class problem with  $K > 2$ , the Fisher criterion is clearly suboptimal. This is seen by a decomposition (Section 2) of the  $K$ -class Fisher criterion into a sum of  $\frac{1}{2}K(K-1)$  two-class criteria, where it becomes obvious that large class distances are overemphasized. The resulting transformation preserves the distances of already well-separated classes, causing a large overlap of neighboring classes, which is suboptimal with respect to the classification rate.

The decomposition, however, allows us to weight the contribution of individual class pairs to the overall criterion in order to improve upon LDA. The weighting scheme discussed in this paper (Section 3) is called the *approximate pairwise accuracy criterion* (aPAC) [10]: Here, the weighting is derived from an attempt to

- M. Loog is with the Image Sciences Institute, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands. E-mail: marco@isi.uu.nl.
- R.P.W. Duin is with the Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands. E-mail: duin@ph.tn.tudelft.nl.
- R. Haeb-Umbach is with Philips Research Laboratories Aachen, Weishausstrasse 2, 52066 Aachen, Germany. E-mail: Reinhold.Haeb@nt.uni-paderborn.de.

Manuscript received 10 May 2000; revised 15 Dec. 2000; accepted 2 Mar. 2001.

Recommended for acceptance by C. Brodley.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112069.

approximate the Bayes error for pairs of classes. While this approach to linear dimension reduction can be viewed as a generalization of LDA, the computational simplicity of LDA is retained: A generalized eigenvalue problem has to be solved and no complex iterative optimization is required.

Section 4 compares the LDR approach based on our aPAC with LDA and with a neural network-based approach. Part of the theory was previously reported in [10] and the experimental results have already been published in [4]. Conclusions are drawn in Section 5.

Several alternative approaches to multiclass LDR are known. In some of them, the problem is stated as an ML estimation task, e.g., [9], [7], in others the divergence is used as a measure for class separation [3]. These criteria, however, also are not directly related to the classification rate. This also holds for the eigenvalue decomposition based approach by Young and Odell [14]. Procedures that deal with the class overlap problem are usually iterative and, thereby, much more computationally demanding, e.g., the Patrick-Fisher approach described in [8], the nonlinear principal component analysis by neural networks [12], and the general, nonparametric approach suggested by Buturovic [1].

### 2 THE FISHER CRITERION AND ITS NONOPTIMALITY

Multiclass LDR is concerned with the search for a linear transformation that reduces the dimension of a given  $n$ -dimensional statistical model, consisting of  $K$  classes, to  $d$  ( $d < n$ ) dimensions, while preserving a maximum amount of discrimination information in the lower-dimensional model. Since it is, however, in general, too complex to use the Bayes error directly as a criterion, one resorts to criteria that are suboptimal but that are easier to optimize. LDA is such a suboptimal approach. A transformation matrix  $\mathbf{L} \in \mathbb{R}^{d \times n}$  is determined which maximizes  $J_F$ , the so-called Fisher criterion:

$$J_F(\mathbf{A}) = \text{tr} \left( (\mathbf{A} \mathbf{S}_W \mathbf{A}^T)^{-1} (\mathbf{A} \mathbf{S}_B \mathbf{A}^T) \right). \quad (1)$$

Here,  $\mathbf{S}_B := \sum_{i=1}^K p_i (\mathbf{m}_i - \bar{\mathbf{m}})(\mathbf{m}_i - \bar{\mathbf{m}})^T$  and  $\mathbf{S}_W := \sum_{i=1}^K p_i \mathbf{S}_i$  are the between-class scatter matrix and the pooled within-class scatter matrix, respectively;  $K$  is the number of classes,  $\mathbf{m}_i$  is the mean vector of class  $i$ ,  $p_i$  is its a priori probability, and the overall mean  $\bar{\mathbf{m}}$  equals  $\sum_{i=1}^K p_i \mathbf{m}_i$ . Furthermore,  $\mathbf{S}_i$  is the within-class covariance matrix of class  $i$ . As can be seen from (1), LDA maximizes the ratio of between-class scatter to average within-class scatter in the lower-dimensional space. The solution to this optimization problem is obtained by an eigenvalue decomposition of  $\mathbf{S}_W^{-1} \mathbf{S}_B$  and taking the rows of  $\mathbf{L}$  to equal the  $d$  eigenvectors corresponding to the  $d$  largest eigenvalues [6]. As long as  $d \geq K-1$ , no information is lost when the classes are normally distributed. Any reduction of dimensionality below  $K-1$  will, however, disturb the class distances. So, now the question arises: How do we find a subspace in which a projection of the class means preserves these distances such that the class separability is maintained as good as possible?

As a part of our approach to this question, the between-class scatter matrix,  $\mathbf{S}_B$ , is rewritten as follows (see [10] for a proof):

$$\mathbf{S}_B = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T. \quad (2)$$

Notice that the decomposition enables us to write the between-class scatter matrix in terms of class-mean differences and that the term  $(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$  is actually the between-class scatter

matrix for the classes  $i$  and  $j$  in a two-class model. Using this decomposition in (1), we obtain for the Fisher criterion

$$J_F(\mathbf{A}) = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \text{tr}((\mathbf{A}\mathbf{S}_W\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{S}_{ij}\mathbf{A}^T)), \quad (3)$$

where  $\mathbf{S}_{ij} := (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$ . Hence, we see that the  $K$ -class Fisher criterion can be decomposed in  $\frac{1}{2}K(K-1)$  two-class Fisher criteria. In this context, we refer to these criteria as *pairwise Fisher criteria*.

Now, to simplify the further discussion, assume the pooled within-class scatter matrix to equal the  $n \times n$ -identity matrix  $\mathbf{I}_n$  and assume that the rows of  $\mathbf{A}$  are scaled to be orthonormal:  $\mathbf{A}\mathbf{A}^T = \mathbf{I}_d$ . These restrictions do not effect the validity of our final conclusions (see Section 3.3). We then obtain the following expression for the Fisher criterion (3):

$$J_F(\mathbf{A}) = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \delta_{ij}^2,$$

where  $\delta_{ij}^2 := \text{tr}(\mathbf{A}\mathbf{S}_{ij}\mathbf{A}^T) = (\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m}_j)^T(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m}_j)$  is the squared distance between the means of class  $i$  and class  $j$  in the dimension reduced model. Hence, we see that the LDA solution to LDR is the linear transformation that maximizes the mean squared distance between the classes in the lower-dimensional space. This, however, is clearly different from minimizing the classification error.

To illustrate that, consider an  $n$ -dimensional model that is to be reduced to one dimension. Assume that one class is located remotely from the other classes and can be considered an *outlier*. In this case, the direction to project on found by optimizing the Fisher criterion is the one that separates the outlier as much from the remaining classes as possible. In maximizing the squared distances, pairs of classes, between which there are large distances, completely dominate the eigenvalue decomposition. As a consequence, there is a large overlap among the remaining classes, leading to an overall low and suboptimal classification rate. Hence, *in general*, LDR by LDA is not optimal with respect to minimizing the classification error rate in the lower-dimensional space. Because outlier classes dominate the eigenvalue decomposition, the LDR transform obtained tends to over-weight the influence of classes that are already well-separated.

### 3 WEIGHTED PAIRWISE FISHER CRITERIA

#### 3.1 Reduction to One Dimension

We now modify the Fisher criterion such that it is more closely related to the classification error. However, we would like to keep the general form of (3) because then the optimization can again be carried out by solving a generalized eigenvalue problem without having to resort to complex iterative optimization schemes. To do so, (3) is generalized by introducing a weighting function  $\omega$ :

$$J_\omega(\mathbf{A}) = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \omega(\Delta_{ij}) \text{tr}((\mathbf{A}\mathbf{S}_W\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{S}_{ij}\mathbf{A}^T)), \quad (4)$$

where  $\omega: \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$  is a weighting function that depends on the Mahanalobis distance

$$\Delta_{ij} := \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)}$$

between the classes  $i$  and  $j$  in the original model. We call these criteria *weighted pairwise Fisher criteria*. This is a reasonable extension because the Bayes error between two classes also depends on the Mahanalobis distance.

Finding a solution  $\mathbf{L}$  that optimizes such a criterion is similar to optimizing the Fisher criterion and comes down to determining an eigenvalue decomposition of the matrix

$$\mathbf{S}_W^{-1} \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \omega(\Delta_{ij}) \mathbf{S}_{ij} \quad (5)$$

and taking the rows of the  $d \times n$ -matrix  $\mathbf{L}$  to equal the  $d$  eigenvectors corresponding to the  $d$  largest eigenvalues. Clearly, choosing  $\omega$  to be the constant function that maps  $\Delta_{ij}$  to 1 results in the ordinary Fisher criterion.

We would like to introduce a weighting function such that the contribution of each class pair depends on the Bayes error rate between the classes. Let us again assume that  $\mathbf{S}_W = \mathbf{I}_n$  and, hence,  $\Delta_{ij}$  equals the ordinary Euclidean distance. (The general case is discussed in Section 3.3.) Then, a mean pairwise accuracy criterion can be stated as follows:

$$J_{\mathcal{A}}(\mathbf{A}) := \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \mathcal{A}_{ij}(\mathbf{A}). \quad (6)$$

Here,  $\mathcal{A}_{ij}(\mathbf{A})$  denotes the accuracy (one minus Bayes error) with respect to the classes  $i$  and  $j$  in the lower-dimensional model obtained by the transformation  $\mathbf{A}$ . Note that this criterion is different from a maximization of the Bayes accuracy of the  $K$ -class problem.

To illustrate what additional approximations we choose to introduce to bring the pairwise accuracy criterion into the form of (4), let us first consider a two-class model which is to be projected onto one dimension. The model is depicted in Fig. 1a. It consists of two normally distributed classes  $i$  and  $j$  having identity covariance matrices and equal a priori probability. The distance between the means is denoted by  $\Delta_{ij} = \|\mathbf{m}_{ij}\| = \|\mathbf{m}_i - \mathbf{m}_j\|$ . Furthermore, the vector  $\mathbf{v}$  denotes the vector we project the model on in going from two dimensions to one. This vector equals  $\mathbf{v} = (\cos \alpha, \sin \alpha)$ , where  $\alpha$  is the angle between  $\mathbf{v}$  and the axis perpendicular to  $\mathbf{m}_{ij}$ .

The accuracy  $\mathcal{A}_{ij}$  in the one-dimensional model obtained after projection onto  $\mathbf{v}$  can be expressed in terms of  $\alpha$  and  $\Delta_{ij}$ :  $\mathcal{A}_{ij}(\mathbf{v}) = \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{\Delta_{ij} \sin \alpha}{2\sqrt{2}}\right)$ , which equals one minus the Bayes error of two normal distributed classes with variance one and distance  $\Delta_{ij} |\sin \alpha|$  between the class means. On the other hand, (4) reads for this particular model

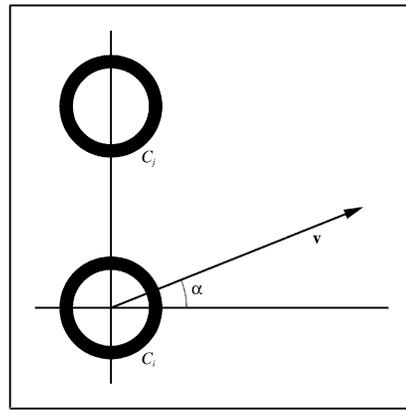
$$J_\omega(\mathbf{v}) = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \omega(\Delta_{ij}) \text{tr}(\mathbf{v}\mathbf{S}_{ij}\mathbf{v}^T) \quad (7)$$

(with  $K=2$ ), where the matrix  $\mathbf{A}$  has been replaced by the row vector  $\mathbf{v}$  since we are reducing the dimension to 1. Note that the two criteria (7) and (6) are *not* equal for all values of  $\alpha$  (see Figs. 1b and 1c).

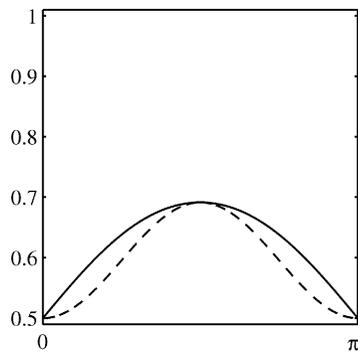
We have chosen to approximate (6) by an expression of the form (7) such that there is equality (up to an additive constant of  $\frac{1}{2}$ ) at the extreme values of  $\alpha$ , i.e., where  $\alpha$  equals 0,  $\pi/2$  and  $\pi$ . This results in the following weighting function:  $\omega(\Delta_{ij}) = \frac{1}{2\Delta_{ij}} \text{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right)$ . To see this, let  $\mathbf{v}$  be a vector in the direction of  $\mathbf{m}_{ij}$ , i.e.,  $\mathbf{v} = \frac{\mathbf{m}_{ij}}{\|\mathbf{m}_{ij}\|}$ . Then, we obtain for the summand in (7),

$$\begin{aligned} \omega(\Delta_{ij}) \text{tr}(\mathbf{v}\mathbf{S}_{ij}\mathbf{v}^T) &= \\ \frac{1}{2\Delta_{ij}} \text{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right) \frac{\mathbf{m}_{ij}^T}{\|\mathbf{m}_{ij}\|} \mathbf{m}_{ij} \mathbf{m}_{ij}^T \frac{\mathbf{m}_{ij}}{\|\mathbf{m}_{ij}\|} &= \frac{1}{2} \text{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right), \end{aligned} \quad (8)$$

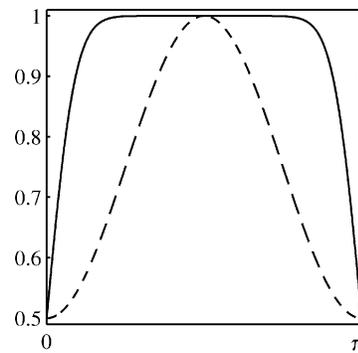
which is exactly the value of the accuracy for  $\alpha = \frac{\pi}{2}$ , up to an additive constant of  $\frac{1}{2}$ . The constant, however, does not influence the optimization. If  $\mathbf{v}$  is perpendicular to  $\mathbf{m}_{ij}$  (i.e.,  $\alpha = 0, \pi$ ), then



(a)



(b)



(c)

Fig. 1. (a) A two-class model in which  $\mathbf{v}$ , determined by the angle  $\alpha$ , is the vector to project on. (b) and (c) Illustration of the Bayes accuracy,  $\mathcal{A}_{ij}(\mathbf{v}) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{\Delta_{ij} \sin \alpha}{2\sqrt{2}}\right)$ , (solid line) and the approximation (dashed line) versus angle  $\alpha$  for the two-class model of (a) for two different values of  $\Delta_{ij}$ : (b)  $\Delta_{ij} = 1$ , (c)  $\Delta_{ij} = 9$ .

$\operatorname{tr}(\mathbf{v}^T \mathbf{S}_{ij} \mathbf{v})$  is zero and again the Bayes accuracy, up to a constant  $\frac{1}{2}$ , results.

In Fig. 1, we illustrate the kind of approximation that is obtained by expression (7) for two values of  $\Delta_{ij}$ ; the graphs of  $\mathcal{A}_{ij}(\mathbf{v})$  (solid curves) and the approximation with a constant of  $\frac{1}{2}$  added (dashed curves) are plotted against the variable  $\alpha$ . We see that the approximation underestimates the accuracy, except in the extrema of the accuracy, where it exactly equals the accuracy.

### 3.2 Reduction to Multiple Dimensions

Unfortunately, (7) only provides an approximation of the mean pairwise accuracy if we reduce the dimension to one since  $\mathbf{v}$  is a vector. In the general case of reducing to  $d$  dimensions with  $d \geq 1$ , we apply a procedure similar to LDA: Determine an eigenvalue decomposition of (5) and take the rows of the LDR transformation  $\mathbf{L}$  to equal the  $d$  eigenvectors corresponding to the  $d$  largest eigenvalues. This matrix  $\mathbf{L}$  maximizes (4). To see that this procedure delivers the equivalent approximation to the pairwise accuracy as the technique described in Section 3.1 did for the reduction to one dimension, consider the following argument.

Because we still assume that  $\mathbf{S}_W = \mathbf{I}_n$ ,  $\mathbf{L}$  is built up of eigenvectors of  $\sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_{ij}$  (N.B., we leave out the limits of the sums when the formulas are in the text), which is a symmetric matrix, hence there are orthogonal eigenvectors, which we denote by  $\mathbf{e}_1$  to  $\mathbf{e}_d$ . Using this, we can write out (4), with  $\mathbf{L} = (\mathbf{e}_1, \dots, \mathbf{e}_d)^T$  substituted for  $\mathbf{A}$ , as follows:

$$J_w(\mathbf{L}) = \sum_{m=1}^d \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \omega(\Delta_{ij}) \operatorname{tr}(\mathbf{e}_m^T \mathbf{S}_{ij} \mathbf{e}_m). \quad (9)$$

Notice that every term  $\sum \sum p_i p_j \omega(\Delta_{ij}) \operatorname{tr}(\mathbf{e}_m^T \mathbf{S}_{ij} \mathbf{e}_m)$  in (9) is similar to (7). We see that maximizing  $J_w$  means that we determine those  $d$  orthogonal directions for which the sum of the accuracies in these directions is maximal. Furthermore, assuming that the eigenvector  $\mathbf{e}_m$  corresponds to the  $m$ th largest eigenvalue, we see that  $\mathbf{e}_1$  is the direction in which the approximate mean pairwise accuracy is maximal. The eigenvector  $\mathbf{e}_2$  is, in fact, also an eigenvector that maximizes this accuracy, however, now under the restriction that its direction is perpendicular to  $\mathbf{e}_1$ . The same holds for  $\mathbf{e}_3$ , which should be perpendicular to  $\mathbf{e}_1$  and  $\mathbf{e}_2$ , etc. Hence, the  $m$ th eigenvector  $\mathbf{e}_m$  determines the direction that attains the maximal approximate mean pairwise accuracy in the  $(n - m + 1)$ -dimensional space that is perpendicular to the space spanned by the eigenvectors  $\mathbf{e}_1$  to  $\mathbf{e}_{m-1}$ .

Our criterion (4), with  $\omega(\Delta_{ij}) = \frac{1}{2\sqrt{2}} \operatorname{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right)$ , approximates the mean accuracy among pairs of classes, hence, we call it an *approximate pairwise accuracy criterion* (aPAC).

### 3.3 General Within-Class Scatter Matrix

In this section, we generalize our findings to a model, where the within-class scatter matrix  $\mathbf{S}_W$  no longer equals the identity matrix.

Applying the linear transform  $\mathbf{S}_W^{-\frac{1}{2}}$  to the original statistical model, we obtain a new statistical model in which the within-class scatter matrix equals the identity matrix. Hence, in this model, distances  $\Delta_{ij}$  come down to the ordinary Euclidean distance

between class means. For a model in which the within-class scatter matrix equals  $\mathbf{I}_n$ , we already argued that  $J_\omega$ , as defined in (9), is a good LDR criterion. Now, let  $\mathbf{L}'$  be a  $d \times n$ -matrix that maximizes  $J_\omega$ . This LDR transform can also be used for reducing the dimension of the original statistical model: We simply use  $\mathbf{L}'\mathbf{S}_W^{-\frac{1}{2}}$  as the LDR transformation, i.e., the original feature vectors are first transformed to the new statistical model by means of  $\mathbf{S}_W^{-\frac{1}{2}}$  and afterward reduced in dimension by  $\mathbf{L}'$ .

We now show that  $\mathbf{L}'\mathbf{S}_W^{-\frac{1}{2}}$  maximizes the weighted pairwise Fisher criterion in the original model. This in turn shows that we can determine and maximize this criterion directly for the original model without explicitly transforming the original model to a new model in which the within-class scatter matrix is  $\mathbf{I}_n$ . To prove the former statement, let  $\mathbf{e}_m$  be an eigenvector of  $\sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}}$ , which is the generalized between-class scatter matrix after applying  $\mathbf{S}_W^{-\frac{1}{2}}$ . Hence,  $\mathbf{L}'$ , which maximizes the weighted pairwise Fisher criterion in this model, is built up of these eigenvectors  $\mathbf{e}_m$ . Now, let  $\lambda_m$  be the eigenvalue associated with  $\mathbf{e}_m$  and notice that the distances  $\Delta_{ij}$  between pairs of class-means are equal for every pair  $(i, j)$  in both models. Regarding the foregoing, we can write the following identities:

$$\begin{aligned} & \left( \mathbf{S}_W^{-1} \sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_{ij} \right) \mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_m \\ &= \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_W^{-\frac{1}{2}} \sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_m \\ &= \mathbf{S}_W^{-\frac{1}{2}} \sum \sum p_i p_j \omega(\Delta_{ij}) \lambda_m \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_m = \mathbf{S}_W^{-\frac{1}{2}} \lambda_m \mathbf{e}_m = \lambda_m \mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_m. \end{aligned}$$

This shows that, if  $\mathbf{e}_m$  is an eigenvector of  $\sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}}$  with associated eigenvalue  $\lambda_m$ , then  $\mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_m$  is an eigenvector of  $\mathbf{S}_W^{-1} \sum \sum p_i p_j \omega(\Delta_{ij}) \mathbf{S}_{ij}$  with the same associated eigenvalue  $\lambda_m$ . This in turn shows that  $\mathbf{L}'\mathbf{S}_W^{-\frac{1}{2}} = (\mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_1, \dots, \mathbf{S}_W^{-\frac{1}{2}} \mathbf{e}_d)^T$  maximizes the weighted pairwise Fisher criterion in the original model because it is built up of the  $d$  eigenvector associated with the  $d$  largest eigenvalues.

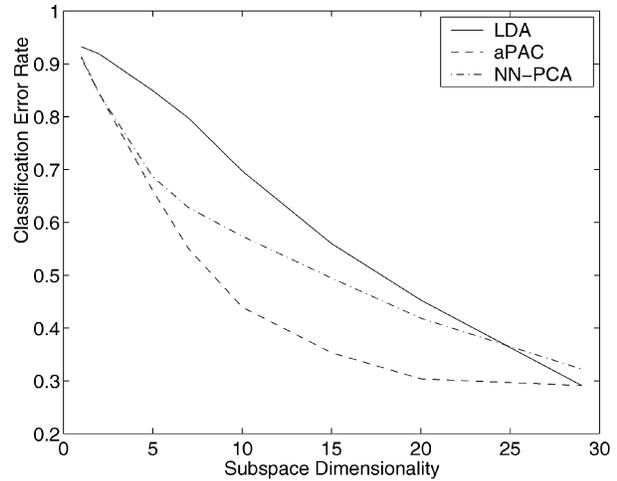
We conclude that an LDR transformation of the form  $\mathbf{L}'\mathbf{S}_W^{-\frac{1}{2}}$  can be found directly in the original model by maximizing  $J_\omega$  as defined in (4), which is done by means of a simple eigenvalue decomposition, as in LDA.

## 4 EXPERIMENTAL RESULTS

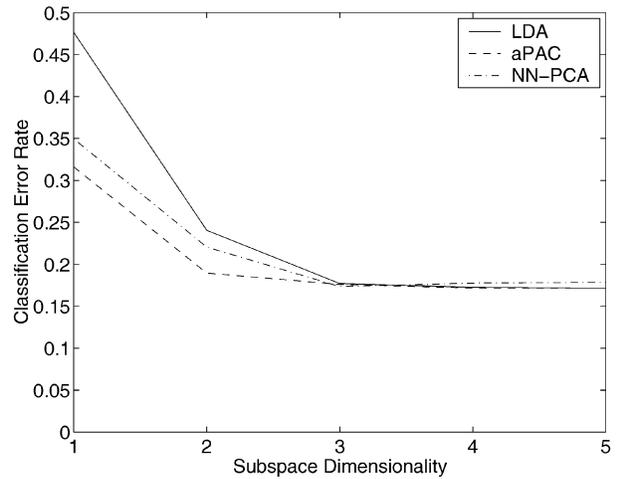
In order to test the LDR transformation obtained by means of the aPAC we performed two experiments, one on artificial data and one on real data. See, also, [4] for a more extensive description of the experiments.

In the artificial problem, we assume known class means and equal class covariance matrices  $\mathbf{I}_n$ . A set of 30 class means is generated from a 30-dimensional normal distribution with covariance matrix  $4\mathbf{I}_n$ . Prior probabilities are assumed to be equal for all classes, i.e.,  $\frac{1}{30}$ . Linearly reduced subspaces with dimensions 1 to 29, for which the Fisher criterion and the aPAC are maximal and which are obtained by projecting the original space with the transformation  $\mathbf{L}$ , are computed. The linear separability of these subspaces is estimated by a Monte Carlo procedure. Classification errors averaged over 10 experiments are shown in Fig. 2a for standard LDA and for the aPAC criterion.

For comparison purposes, Fig. 2a also includes results obtained by an  $[n, d, n]$  autoassociative neural network (denoted NN-PCA). It has linear neurons in the hidden layer, which thereby constructs the linear subspace, and a sigmoidal transfer function at the



(a)



(b)

Fig. 2. (a) Classification error rate as a function of subspace dimensionality for artificial problem with 30 classes and 30 features. (b) Classification error rate as a function of subspace dimensionality for Landsat data set.

output, see [12]. As can be seen in Fig. 2a, the neural network procedure performs similarly as the aPAC approach for low-dimensionalities. For larger dimensionalities, however, it performs significantly worse in our implementation (we used Matlab's Neural Network Toolbox). Moreover, it needs significantly more computational effort.

For the real data set, we used the Landsat data set as used in the Statlog project [11]. This is a 6-class problem with 36 features. It consists of a total of 6,435 objects, of which 4,435 are in the training set and the remaining 2,000 are in the test set. The three methods are used again for finding subspaces of one to five dimensions, this time including the prewhitening step as means and variances are unknown. For each subspace, a linear classifier assuming normal densities is computed on the training data used for obtaining the mapping. Resulting test errors, shown in Fig. 2b, confirm the results of the artificial problem: The nonlinear criteria improve standard LDA with the aPAC being superior to the neural network.

## 5 CONCLUSIONS

In this paper, we proposed a new class of computationally inexpensive LDR criteria which generalize the well-known Fisher criterion used in LDA. Noting that the  $K$ -class Fisher criterion can be decomposed into  $\frac{1}{2}K(K-1)$  two-class Fisher criteria, the generalization is obtained by introducing a weighting of the contributions of individual class pairs to the overall criterion.

An important property of the criteria we presented here is that they can be designed to confine the influence of outlier classes on the final LDR transformation. This makes them more robust than LDA. However, it cannot be guaranteed that the new criteria always lead to improved classification rate because of the various approximations we had to introduce to arrive at a solution which is computationally simple.

An interesting subclass of these criteria are the *approximate pairwise accuracy criteria* (aPAC). These aPAC take into account classification errors occurring between pairs of classes, unlike the Fisher criterion that is merely based on measures of spread. In the article, we investigated one particular aPAC and compared its performance to that of LDA in two experiments, one on artificial data and one on real-world data. The experiments clearly showed the improvement that is possible when utilizing aPAC instead of the Fisher criterion.

## REFERENCES

- [1] L.J. Buturovic, "Toward Bayes-Optimal Linear Dimension Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 420-424, 1994.
- [2] N.A. Campbell, "Canonical Variate Analysis—A General Model Formulation," *Australian J. Statistics*, vol. 26, pp. 86-96, 1984.
- [3] H.P. Decell and S.M. Mayekar, "Feature Combinations and the Divergence Criterion," *Computing and Math. with Application*, vol. 3, pp. 71-76, 1977.
- [4] R.P.W. Duin, M. Loog, and R. Haeb-Umbach, "Multi-Class Linear Feature Extraction by Nonlinear PCA," *Proc. Int'l Conf. Pattern Recognition*, 2000.
- [5] R.A. Fisher, "The Statistical Utilization of Multiple Measurements," *Ann. Eugenics*, vol. 8 pp. 376-386, 1938.
- [6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
- [7] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixtures." *J. Royal Statistical Soc., B*, vol. 58, pp. 155-176, 1996.
- [8] J. Kittler, "Feature Selection and Extraction," *Handbook of Pattern Recognition and Image Processing*. Academic Press, 1986.
- [9] N. Kumar and A.G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Comm.*, vol. 26, pp. 283-297, 1998.
- [10] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion*. Delft Univ. Press, 1999.
- [11] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.
- [12] E. Oja, "The Nonlinear PCA Learning Rule in Independent Component Analysis," *Neurocomputing*, vol. 17, pp. 25-45, 1997.
- [13] C.R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," *J. Royal Statistical Soc., B*, vol. 10, pp. 159-203, 1948.
- [14] D.M. Young and P.L. Odell, "A Formulation and Comparison of Two Linear Feature Selection Techniques Applicable to Statistical Classification," *Pattern Recognition*, vol. 17, pp. 331-337, 1984.

► For further information on this or any computing topic, please visit our Digital Library at <http://computer.org/publications/dlib>.