

# MULTI-CLASS LINEAR DIMENSION REDUCTION BY GENERALIZED FISHER CRITERIA

Marco Loog<sup>2,1</sup>

Reinhold Haeb-Umbach<sup>1</sup>

<sup>1</sup>Philips Research Laboratories Aachen  
Weissshausstrasse 2, 52066 Aachen, Germany  
Reinhold.Haeb@philips.com

<sup>2</sup>SSOR, Faculty of Information Technology and Systems  
Delft University of Technology, The Netherlands

## ABSTRACT

Linear Discriminant Analysis is in general unable to find the lower-dimensional feature space which maximizes the class discrimination, even if the class distributions can be assumed to be very simple, e.g. Gaussians with identical covariance matrices. In this paper we reformulate the  $K$ -class Fisher criterion as a sum of  $K(K-1)/2$  2-class Fisher criteria. This formulation allows to weigh class pair contributions according to their relevance for classification. Further it offers an obvious way how to cope with heteroscedastic models. We propose a particular weighting scheme which attempts to approximate the pairwise Bayes error. Moderate improvements are obtained on the TIMIT phoneme classification task.

## 1. INTRODUCTION

Reducing the feature vector dimension of a statistical pattern classifier is desirable for two reasons. First, it results in lower computational and memory demands of the system. Second, and more important, it decreases the number of model parameters to be estimated, which in general will lead to more accurate estimates given a limited training database.

In this paper we are concerned with *linear* dimension reduction, although the objective function may well be nonlinear. Fisher has introduced a technique of dimension reduction of a two-class problem to a one-dimensional subspace. This technique has later been extended to handle  $K$ -class problems with  $K > 2$ , and it has become known as "Linear Discriminant Analysis" (LDA). Although the objective function of LDA is related to class discrimination, we will see in this paper that it will in general not find the lower-dimensional feature space with minimum classification error rate. We show that the transformation tries to preserve distances of already well separated classes, which may result in a large overlap or even occlusion of neighboring classes.

Here we propose a generalization of the Fisher Criterion, which allows to deemphasize the contributions of classes, which are far apart from each other. It also allows to take into account differences in class covariances, thus being an extension of LDA towards heteroscedastic data. At the same time the transformation matrix is still computed via a generalized eigenvalue problem. No iterative optimization is required.

The application of LDA to speech recognition has shown consistent gains for several recognition tasks, despite its simplicity and obvious simplifying model assumptions. The hope that

there is more to gain by finding more sophisticated multi-class dimension reduction algorithms has therefore spurred a lot of research in optimal dimension reduction, heteroscedastic models and decorrelation.

One category of approaches takes the viewpoint of Campbell [1]. He has shown that the determination of the LDA transform is equivalent to finding the Maximum Likelihood parameter estimates of a Gaussian model which assumes 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 (so-called "homoscedastic model"). Kumar and Andreou extended LDA to "heteroscedastic discriminant analysis (HDA), where the within-class covariances need no longer be equal [6]. Gopinath [4] developed a theory of "Constrained Maximum Likelihood" estimation, of which a model which assumes equal class covariances is an example of such a constrained model. If LDA or HDA is followed by a diagonalizing linear transform, significant improvements in classification accuracy have been obtained [8]. Hastie has developed mixture discriminant analysis, where the class distributions are assumed to be mixtures of Gaussians rather than unimodal Gaussians [5]. The assumptions of heteroscedastic and mixture models better fit with the typical modeling of speech data in Hidden Markov Model based speech recognizers.

Other approaches modify the objective function or the measure of class distance to arrive at linear dimension reductions which have more discriminative power. Demuynck [2] uses a minimum divergence criterion between posterior class distributions in the original and transformed space to estimate an HDA matrix. Thomae [9] iteratively optimizes the LDA matrix in order to maximize the difference in distance between the observation and the correct class mean and between the observation and the closest competing class mean. This is done by employing the generalized probabilistic descend algorithm known from discriminative training.

In Section 2 we reformulate the  $K$ -class Fisher criterion as a sum of  $\frac{1}{2}K(K-1)$  2-class Fisher criteria and demonstrate that the Fisher criterion is often unable to find the reduced space with minimum classification error. This formulation allows to control the contributions of individual class pairs to the overall objective function according to their contribution to the classification rate. In Section 3 we propose a weighting which is derived from the classification error, both for homoscedastic and heteroscedastic models. Finally, Section 4 presents the results of phoneme recognition experiments conducted on the TIMIT database.

## 2. FISHER CRITERION

Multi-class Linear Dimension Reduction (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  $d$  ( $d \leq n$ ) dimensions. The transformation should be such that a maximum amount of discrimination information is preserved 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. Let each class  $i$ ,  $1 \leq i \leq K$ , be characterized by its mean  $\mathbf{m}_i$ , covariance  $\mathbf{S}_i$ , and a priori probability  $p_i$ . In LDA the class information is condensed in two scatter matrices, the between-class scatter

$$\mathbf{S}_B := \sum_{i=1}^K p_i (\mathbf{m}_i - \bar{\mathbf{m}}) (\mathbf{m}_i - \bar{\mathbf{m}})^T, \quad (1)$$

where  $\bar{\mathbf{m}} = \sum_{i=1}^K p_i \mathbf{m}_i$  denotes the overall mean, and the average within-class scatter

$$\mathbf{S}_W := \sum_{i=1}^K p_i \mathbf{S}_i. \quad (2)$$

The goal of LDA is to find a linear transformation  $f: \mathbb{R}^n \rightarrow \mathbb{R}^d$ ,  $y = f(x) = \mathbf{L}x$ , with  $\mathbf{L}$  a  $d \times n$  matrix of rank  $d$  such that the so-called Fisher Criterion is maximized:

$$J_F(\mathbf{L}) = \text{tr}((\mathbf{L}\mathbf{S}_W\mathbf{L}^T)^{-1}(\mathbf{L}\mathbf{S}_B\mathbf{L}^T)), \quad (3)$$

where  $\text{tr}(\cdot)$  denotes the trace of a matrix. After a pre-whitening step, which transforms  $\mathbf{S}_W$  into the identity matrix, the problem is completely defined by the set of  $K$  class means. In [7] it is shown that the between scatter matrix can be written as a sum of pairwise distances as follows:

$$\mathbf{S}_B = \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \mathbf{D}_{ij}. \quad (4)$$

where

$$\mathbf{D}_{ij} := (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \quad (5)$$

is the outer product of the difference vector of the means of the classes  $i$  and  $j$ . It captures the optimal direction in which the two classes can be separated, and is called ‘‘distance matrix’’ in the following.

Using the notation  $\mathbf{m}_{ij} = (\mathbf{m}_i - \mathbf{m}_j)$  and  $\delta_{ij} = \|\mathbf{m}_{ij}\|$  it is readily seen that

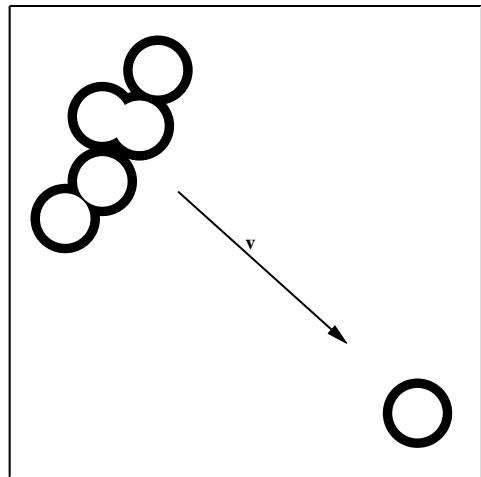
$$\mathbf{D}_{ij} \mathbf{m}_{ij} = \delta_{ij}^2 \mathbf{m}_{ij}, \quad (6)$$

i.e.  $\mathbf{m}_{ij}$  is an eigenvector of  $\mathbf{D}_{ij}$  with eigenvalue  $\delta_{ij}^2$ . In the Fisher Criterion eq. (3) the trace is computed. Note that  $\text{tr}(\mathbf{D}_{ij}) = \delta_{ij}^2$ . Thus, the larger  $\delta_{ij}^2$  the more the direction  $\mathbf{m}_{ij}$  becomes visible in the eigenvectors corresponding to the larger eigenvalues of  $\mathbf{S}_W^{-1} \mathbf{S}_B$ , i.e. the more they determine the transformation  $\mathbf{L}$ .

As an example, consider the two-dimensional 6-class model depicted in Fig. 1. Each class is represented by a circle of the

same radius, indicating that the within-class scatter matrix is assumed to be the identity matrix and that the a priori probabilities are equal for each class. Let  $j_0$  denote the class in the lower right corner of the figure. Provided that this class is sufficiently far apart from the other classes, the contributions  $\delta_{ij_0}^2$ ,  $1 \leq i \leq K, i \neq j_0$ , will dominate the between-class scatter and thus the Fisher criterion. As a consequence the direction indicated by the arrow  $\mathbf{v}$  will be selected as principal discriminant. However, this direction maps the classes  $i, i \neq j_0$  into one cluster with a lot of overlap between the classes, entailing many classification errors in the one-dimensional space. From a classification point of view the direction orthogonal to  $\mathbf{v}$  would have been more favorable.

From this example we conclude that for  $K > 2$ , LDR by LDA is not optimal with respect to minimizing the classification error rate in the lower-dimensional space.



**Figure 1:** A two-dimensional 6-class model consisting of five classes arranged in a cluster and one ‘outlier’ class. The vector  $\mathbf{v}$  indicates the direction obtained by LDA onto which the model is projected when the dimension is reduced to one.

## 3. GENERALIZED FISHER CRITERIA

### 3.1. Homoscedastic Model

The major advantage of the above formulation (eq. (4)) is that contributions of individual class pairs to the between-class scatter become visible.

We have seen, that in the LDA case the contributions depend only on the differences of class means and that the contributions are proportional to the square of the distance between the class means.

A first generalization is thus obtained by introducing a weighting function  $\omega(\delta_{ij})$  and replacing  $\mathbf{D}_{ij}$  by

$$\phi(\mathbf{D}_{ij}) := \frac{1}{\delta_{ij}^2} \omega(\delta_{ij}) \mathbf{D}_{ij}. \quad (7)$$

This weighting is used to define a generalized between class scat-

ter matrix

$$\mathbf{S}_{\phi(\mathbf{D})} := \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \phi(\mathbf{D}_{ij}) \quad (8)$$

which replaces  $\mathbf{S}_B$  in the Fisher criterion of eq. (3).

For LDA,  $\phi(\mathbf{D}_{ij}) = \mathbf{D}_{ij}$ , i.e.  $\omega(\delta_{ij}) = \delta_{ij}^2$ . Note that if one chooses  $\omega(\delta_{ij}) = 1$ , then all directions  $\mathbf{m}_{ij}$ ,  $1 \leq i, j \leq K, i \neq j$  have an equal influence on the final criterion (assuming equal a priori probabilities).

Now consider a more sophisticated weighting. The ideal objective function for dimension reduction is the Bayes error. We now would like to approximate this objective function such that it obtains the form of eq. (3), however with the generalized between-class scatter  $\mathbf{S}_{\phi(\mathbf{D})}$  instead of  $\mathbf{S}_B$ . If the form of eq. (3) is retained, the optimum transformation matrix is still obtained as the solution of the generalized eigenvalue problem.

We need the following approximations: First we approximate the  $K$ -class Bayes error by the sum of the 2-class errors, which is an upper bound to the  $K$ -class error. Actually, we employ the accuracy  $\mathcal{A}_{ij}$ , i.e. one minus the Bayes error and define a mean pairwise accuracy criterion as follows:

$$J_{\mathcal{A}}(\mathbf{L}) := \sum_{i=1}^{K-1} \sum_{j=i+1}^K p_i p_j \mathcal{A}_{ij}(\mathbf{L}). \quad (9)$$

Second, the pairwise accuracy  $\mathcal{A}_{ij}$  is approximated such that eq. (9) has the same form as (3) with  $\mathbf{S}_B$  replaced by  $\mathbf{S}_{\phi(\mathbf{D})}$ . Then the optimization can again be carried out by solving a generalised eigenvalue problem, just as in the LDA case.

The accuracy of a Gaussian two-class model with classes  $i$  and  $j$  is a function of the (one-dimensional) line  $\mathbf{v}$  the model is projected on:

$$\mathcal{A}_{ij}(\mathbf{v}) = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left( \frac{\delta_{ij} |\cos \alpha|}{2\sqrt{2}} \right), \quad (10)$$

where  $\delta_{ij} = \|\mathbf{m}_{ij}\|$  is the distance between the class means and  $\alpha$  is the angle between  $\mathbf{m}_{ij}$  and  $\mathbf{v}$ .  $\operatorname{erf}(x)$  is the error function defined as  $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ . In [7] it is shown that the desired approximation is obtained by using the following weighting function in eq. (7):

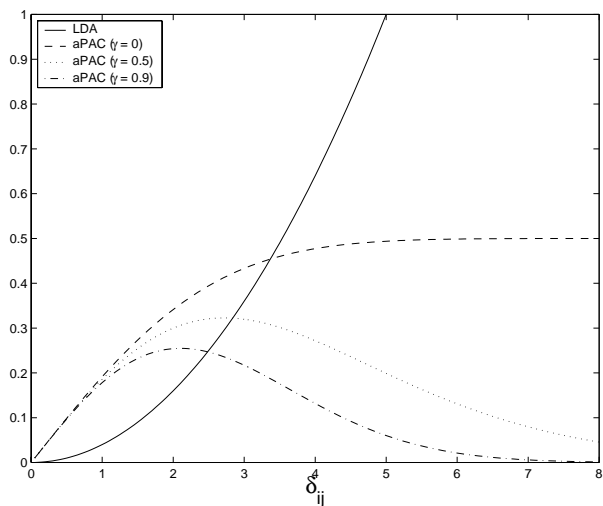
$$\omega(\delta_{ij}) = \frac{1}{2(1-\gamma)} \left( \operatorname{erf} \left( \frac{\delta_{ij}}{2\sqrt{2}} \right) - \operatorname{erf} \left( \frac{\gamma \delta_{ij}}{2\sqrt{2}} \right) \right), \quad (11)$$

$\gamma$  is a parameter which controls the approximation. It should be set somewhere in the range  $[0, 1]$ . We call these kind of criteria ‘‘approximate pairwise accuracy (aPAC)’’ criteria.

Figure 2 compares the LDA weighting with the weighting according to equation (11) for different values of the control parameter  $\gamma$ . It can be seen that, compared to LDA, the influence of already well separated classes on the criterion is reduced.

### 3.2. Heteroscedastic Model

The distance matrices used sofar, eq. (5), only capture the distances between the means of the models. This is not changed by



**Figure 2:** Weighting  $\omega(\delta_{ij})$  as a function of distance  $\delta_{ij}$  for LDA and variants of aPAC. (for LDA  $\delta_{ij}^2$  is multiplied by  $\frac{1}{25}$  for better illustration).

the weighting introduced in the last section. Since the trace of the distance matrix  $\operatorname{tr}(\mathbf{D}_{ij})$  equals the Euclidian distance between  $\mathbf{m}_i$  and  $\mathbf{m}_j$ , we use the subscript  $E$  and denote the distance matrix as

$$\mathcal{D}_E(\mathbf{m}_i, \mathbf{m}_j) := \mathbf{D}_{ij} = (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \quad (12)$$

If the data are not homoscedastic, the Euclidian distance might not be appropriate. Then a measure should be employed which takes also into account differences in the class covariances. Examples are the Mahalanobis distance, the Bhattacharyya distance or the Kulback-Leibler divergence. For example in the case of the Mahalanobis distance, the following distance matrix is defined:

$$\mathcal{D}_M(\mathbf{m}_i, \mathbf{m}_j, \mathbf{S}_i, \mathbf{S}_j) = \mathbf{S}^{-1/2}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1/2} \quad (13)$$

where  $\mathbf{S} = \frac{1}{2}(\mathbf{S}_i + \mathbf{S}_j)$ .

It is straightforward to show that the trace of  $\mathcal{D}_M(\mathbf{m}_i, \mathbf{m}_j, \mathbf{S}_i, \mathbf{S}_j)$  corresponds to the Mahalanobis distance between two normally distributed densities with mean and covariance  $\mathbf{m}_i, \mathbf{S}_i$  and  $\mathbf{m}_j, \mathbf{S}_j$ , respectively:

$$\operatorname{tr}(\mathcal{D}_M) = (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}^{-1}(\mathbf{m}_i - \mathbf{m}_j). \quad (14)$$

Because the trace of  $\mathcal{D}_E(\mathbf{m}_i, \mathbf{m}_j)$  is equal to the eigenvalue  $\delta_{ij}^2$ , see comment after eq. (6), it was possible to achieve a weighting of the contribution of each class pair to the Fisher Criterion by a multiplicative weighting function applied to the distance matrix, eq. (7). However,  $\mathcal{D}_M(\mathbf{m}_i, \mathbf{m}_j, \mathbf{S}_i, \mathbf{S}_j)$  will in general have more than one non-zero eigenvalue. As a consequence, the desired weighting of the eigenvalues can no longer be achieved by a multiplication of the distance matrix. Instead, a function is defined which manipulates the eigenvalues of a matrix. Specifically, we define a function  $f(\mathbf{A})$  of a positive semidefinite matrix argument  $\mathbf{A}$  such that the function is applied to the eigenvalues of  $\mathbf{A}$ , i.e. if  $\mathbf{A} = \mathbf{R}\mathbf{V}\mathbf{R}^{-1}$  is the eigenvalue

decomposition of  $\mathbf{A}$ , with  $\mathbf{V} = \text{diag}(v_1, \dots, v_n)$ , we get

$$f(\mathbf{A}) := \mathbf{R}f(\mathbf{V})\mathbf{R}^{-1} := \mathbf{R}\text{diag}(f(v_1), \dots, f(v_n))\mathbf{R}^{-1}. \quad (15)$$

The function  $\phi$  we choose to approximate the mean pairwise accuracy is defined as follows:

$$\begin{aligned} \phi(\mathcal{D}_M) := & \left( \frac{1}{2} - \frac{\gamma}{2(1-\gamma)} \text{erf} \left( \frac{\sqrt{\lambda_{(1)}^{\mathcal{D}_M}}}{2\sqrt{2}} \right) \right) \mathbf{I} \\ & + \frac{1}{2(1-\gamma)} \text{erf} \left( \frac{((1-\gamma)\sqrt{\mathcal{D}_M} + \gamma\sqrt{\lambda_{(1)}^{\mathcal{D}_M}} \mathbf{I})}{2\sqrt{2}} \right), \end{aligned} \quad (16)$$

$\lambda_{(1)}^{\mathcal{D}_M}$  is the largest eigenvalue of  $\mathcal{D}_M$ . In [7] we show that this function reduces to the eigenvalue weighting of eq. (11) in case the model is homoscedastic.

## 4. EXPERIMENTAL RESULTS

After verifying that the presented approach worked well on some classical pattern classification tasks [3] we conducted phoneme recognition experiments on the TIMIT speech database. The word insertion penalty was adjusted to obtain phone insertion rates in the range of 10% for the sake of comparability with results published by others. We present context-independent phoneme classification results *without* the use of a phone bigram.

The baseline system has 115 context-independent states and 7000 Gaussian mixture components with diagonal covariance matrices. We considered each state as a separate class (i.e.  $K = 115$ ). The frame feature vector consists of 16 static cepstral coefficients, 16 deltas and the second-order derivative of the frame energy. Every 3 consecutive 33-dimensional cepstral vectors were spliced together forming 99-dimensional feature vectors, on which the linear dimension reduction experiments were conducted.

Table 4 shows the results for a dimension reduction of the model from 99 to 35 dimensions. The reduction was done by means of LDA and certain aPAC approaches. For the aPAC experiments we used the Mahalanobis distance matrix, eq. (13) with different values for the approximation control parameter  $\gamma$ . It is important to note that for the aPAC approaches it is not guaranteed that the between-class and pooled within-class covariance matrices are diagonal. After dimension reduction we therefore did an additional linear transformation without dimension reduction, which was either a

- *homoscedastic decorrelation*: simultaneous diagonalization of between-class and average within-class covariance matrix (an LDA transformation without dimension reduction).
- *heteroscedastic decorrelation*: a rotation of the 35-dimensional space such that the 115 within-class covariance matrices are as diagonal as possible [6] (called “maximum-likelihood linear transform in [4]).

As can be seen from the table, aPAC slightly outperforms LDA both for subsequent homoscedastic and heteroscedastic decorrelation. Note that for dimension reduction by LDA of course no homoscedastic decorrelation is required (first result line in table).

**Table 1:** Phone accuracy [%] on TIMIT for different dimension reduction criteria and subsequent decorrelation.

LDR approach	phone accuracy in %
LDA	homosc. decorrelation
Mahalanobis based aPAC ( $\gamma = 0$ )	70.53
Mahalanobis based aPAC ( $\gamma = .5$ )	70.80
	70.80
LDA	heterosc. decorrelation
Mahalanobis based aPAC ( $\gamma = 0$ )	71.83
Mahalanobis based aPAC ( $\gamma = .5$ )	72.11
	72.43

## 5. SUMMARY

In this paper we proposed a generalization of the Fisher criterion by introducing a weighting function which controls the contribution of class pairs. A particular function is presented which weighs contributions of class pairs according to an approximation to their two-class Bayes error rate. The function also allows to take into account heteroscedasticity of the data. Although the weighting is still not optimal as only two-class relations are considered and as the Bayes error can only be approximated, it solves the outlier problem, in that remote classes can no longer dominate the between-class scatter. Still, it retains the computational simplicity of LDA and does not require an iterative optimization procedure as is typical for other dimension reduction schemes.

We tested the novel approach on the TIMIT phoneme classification task and obtained moderate improvements in phoneme accuracy.

## 6. REFERENCES

1. N. A. Campbell. Canonical variate analysis—a general model formulation. *Australian journal of statistics*, 26:86–96, 1984.
2. K. Demuyne, J. Duchateau, and D. Van Campennolle. Optimal feature sub-space selection based on discriminant analysis. In *Proc. EUROSPEECH’99*, pages 1311–1314, 1999.
3. R.P.W. Duin, M. Loog, and R. Haeb-Umbach. Multi-class feature extraction by nonlinear PCA. In *Proc. Int. Conference on Pattern Recognition*, 2000.
4. R. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *Proc. ICASSP’98*, pages 661–664, 1998.
5. T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society series B*, 58:155–176, 1996.
6. N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26:283–297, 1998.
7. M. Loog. Approximate pairwise accuracy criteria for multiclass linear dimension reduction – generalisations of the Fisher criterion. *WBBM Report Series 44*, 1999.
8. R. Sao, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proc. ICASSP’00*, pages 1129–1132, 2000.
9. M. Thomae, G. Ruske, and T. Pfau. A new approach to discriminative feature extraction using model transformation. In *Proc. ICASSP’00*, pages 1615–1618, 2000.